applied sciences

MDPI

Article

Gaze-Dependent Image Re-Ranking Technique for Enhancing Content-Based Image Retrieval

Yuhu Feng [1], Keisuke Maeda [2], Takahiro Ogawa [2] and Miki Haseyama [2,*]

[1] Graduate School of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-ku, Sapporo 060-0814, Hokkaido, Japan; feng@lmd.ist.hokudai.ac.jp
[2] Faculty of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-ku, Sapporo 060-0814, Hokkaido, Japan; maeda@lmd.ist.hokudai.ac.jp (K.M.); ogawa@lmd.ist.hokudai.ac.jp (T.O.)
* Correspondence: mhaseyama@lmd.ist.hokudai.ac.jp

Abstract: Content-based image retrieval (CBIR) aims to find desired images similar to the image input by the user, and it is extensively used in the real world. Conventional CBIR methods do not consider user preferences since they only determine retrieval results by referring to the degree of resemblance or likeness between the query and potential candidate images. Because of the above reason, a "semantic gap" appears, as the model may not accurately understand the potential intention that a user has included in the query image. In this article, we propose a re-ranking method for CBIR that considers a user's gaze trace as interactive information to help the model predict the user's inherent attention. The proposed method uses the user's gaze trace corresponding to the image obtained from the initial retrieval as the user's preference information. We introduce image captioning to effectively express the relationship between images and gaze information by generating image captions based on the gaze trace. As a result, we can transform the coordinate data into a text format and explicitly express the semantic information of the images. Finally, image retrieval is performed again using the generated gaze-dependent image captions to obtain images that align more accurately with the user's preferences or interests. The experimental results on an open image dataset with corresponding gaze traces and human-generated descriptions demonstrate the efficacy or efficiency of the proposed method. Our method considers visual information as the user's feedback to achieve user-oriented image retrieval.

Keywords: re-ranking; gaze trace; content-based image retrieval; image captioning

Citation: Feng, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. Gaze-Dependent Image Re-Ranking Technique for Enhancing Content-Based Image Retrieval. *Appl. Sci.* **2023**, *13*, 5948. https://doi.org/10.3390/app13105948

Academic Editors: Bin Fan and Wenqi Ren

Received: 23 March 2023
Revised: 30 April 2023
Accepted: 8 May 2023
Published: 11 May 2023

1. Introduction

With the popularization of personal electronic terminals such as smartphones and tablets, the amount of visual data available on the internet has been overgrown in recent years [1]. Correspondingly, the performance of related image retrieval technologies is also improving and becoming more sophisticated [2]. Content-based image retrieval (CBIR) is a type of search technology that aims to find images similar to the image input by users (query image), and it is extensively applied in the real world [3]. For example, search engines of image retrieval functions (e.g., Google (http://images.google.it, accessed on 23 February 2023)) and a similar production search of online shopping (e.g., Amazon (https://www.amazon.co.jp/, accessed on 23 February 2023) and eBay (https://www.ebay.com/, accessed on 23 February 2023))are examples of such applications. Conventional CBIR methods have exhibited remarkable precision in retrieving images that bear similarities to the query image, as documented in several studies [4–7]. It is feasible to efficiently and effectively retrieve associated images from a vast database using a single input image. In the last decade, extensive research efforts have been devoted to developing novel theories and models for CBIR, establishing several practical algorithms [8]. However, as the volume of visual data on the web continues to increase, there is a growing imperative to consider

users' subjective preferences during image retrieval to enhance the value of the retrieved data and satisfy the increasing demands of users.

For users, whether the retrieval result is appropriate cannot be judged solely by the image content, but also based on the user's preferences. The aforementioned problem is a crucial challenge known as the semantic gap [9], which CBIR has been facing for a long time. Since the query image input by a user is rich in detail and information, it can be challenging for the CBIR model to accurately localize which specific part of the image the user intends to retrieve a similar one from the dataset [10]. Conventional CBIR methods compare the features of the query image with the features of images in the dataset (candidate images) and rank these candidate images according to their similarity to the query image. Fadaei et al. [11] proposed an approach to extract dominant color descriptors for CBIR. The approach uses a weighting mechanism that assigns more importance to the informative pixels of an image using suitable masks. However, these methods may fail to rank images fitting to the user's preferences with a high rank consistently. To achieve user-oriented image retrieval, one of the possible strategies is to re-order the images of the initial retrieval, satisfying the user's preferences with a higher position.

Re-ranking is an approach to re-order the results of initial retrieval using reliable information and typically plays a role in the post-processing step in image retrieval tasks [12]. Re-ranking methods can be classified into two categories based on the information source employed: self-re-ranking and interactive re-ranking [13]. Self-re-ranking approaches aim to improve the accuracy of initial retrieval results by identifying relevant visual patterns from them and re-ordering them based on external information, such as text labels or class information. For example, Zhang et al. [14] proposed a new method based on a global model for extracting features from the entire image and a local model for extracting features from individual regions of interest in the image. The results from these two models are then combined using a re-ranking approach, which significantly improves the accuracy of the retrieval results. Conversely, interactive re-ranking approaches use user feedback from the initial retrieval results as preference information for re-ordering [13]. Therefore, interactive re-ranking is expected to achieve higher user satisfaction by utilizing feedback reflecting users' preferences.

Recent studies [15–17] have shown that gaze information, which consists of human eye movements, plays a critical role in visual recognition during daily life and involves attention shifts. Gaze information is integral to non-verbal communication in an interaction process between humans and the natural world and is closely related to user's preferences when viewing images [18]. Therefore, by introducing the user's gaze information into the initial retrieval re-order as the user's feedback information for re-ranking, the users are expected to find their desired images in a higher rank. However, directly using gaze trace data in a coordinate format may not accurately capture the relationships between objects in an image, which could be problematic.

To tackle the abovementioned issue, we propose a novel CBIR method utilizing image captioning as a gaze-dependent image re-ranking method. Figure 1 illustrates the underlying concept of our proposed image re-ranking method. Our method leverages the interdisciplinary technology of image captioning. This technique enables machines to automatically produce natural language descriptions for any given image [19]. The proposed method entails developing a neural network that integrates images and gaze information to generate image captioning controlled by gaze traces. The transformer is used in the proposed method, focusing on its characteristics. In contrast to convolutional neural networks (CNNs) that rely on local connectivity, the transformer achieves a global representation through shallow layers and preserves spatial information [20]. Given the expectation that the transformer model is more closely aligned with human visual characteristics than CNNs [21], the proposed method trains a connected caption and gaze trace (CGT) model [22] on the basis of the transformer architecture. This enables the model to learn the intricate relationship between images, captioning, and gaze traces.

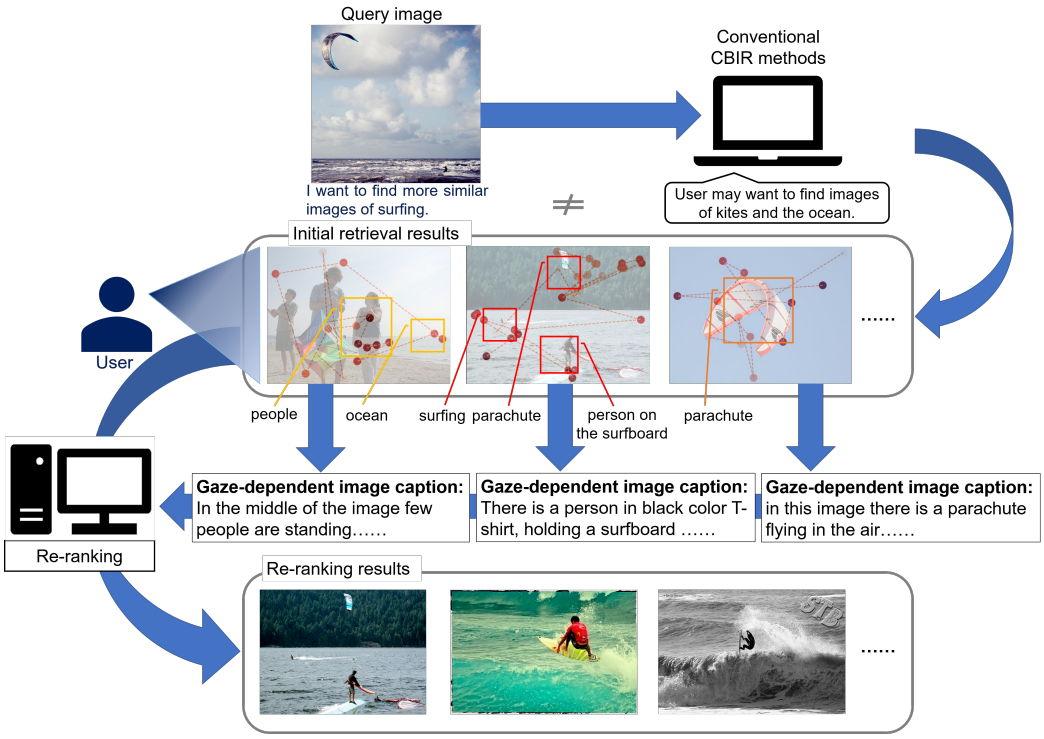


Figure 1. Concept of our method to improve conventional CBIR methods. Conventional methods face the problem of accurately localizing which specific part of the image the user intends to retrieve. Our method attempts to present images that meet the user's favor at a higher rank using the gaze-dependent image caption.

With the introduction of image captioning, which connects image features and gaze information, the proposed method can explicitly express semantic information in images (e.g., relationships between objects that users focus on) to realize the re-ranking that accurately reflects the user's preferences. Specifically, our method uses the Contrastive Language-Image Pre-Training (CLIP) model [23], extensively acknowledged as one of the most sophisticated cross-modal embedding techniques currently available. The primary objective of CLIP is to create a common latent space for computing the similarity between images and text. Therefore, we can compare the gaze-dependent image captions and the candidate images by embedding them in the latent space and ranking the images that reflect the user's preferences higher in the re-order process. Extensive experimentation demonstrates the remarkable performance of the proposed method in re-ranking for image retrieval on a publicly available dataset [24] of annotated images in the MSCOCO dataset [25].

In conclusion, the key contributions of our study are as follows.

- We propose a novel gaze-dependent re-ranking method for CBIR to tackle the "semantic gap" challenge of the conventional CBIR methods.
- We introduce the gaze-dependent image caption to convert coordinate-format visual information into text-format semantic information, thereby realizing re-ranking according to the users' preferences.

Note that we have previously presented some preliminary results of our current work in a prior study [22], where we demonstrated the effectiveness of incorporating gaze-dependent image captioning for achieving personalized image retrieval. In this study, we build upon our previous work and extend it in the following ways. First, we expand on our previous study by utilizing gaze-dependent image captioning as auxiliary information to achieve user-oriented image re-ranking. Second, we evaluate the effectiveness of previous cross-modal retrieval methods and interactive re-ranking methods to validate the robustness of our proposed method. Finally, in our ablation study, we verify the novelty of our method by comparing the effectiveness of incorporating gaze data directly for re-ranking

and transforming the data from coordinate data into a text format to bridge the semantic gap in CBIR effectively.

The remainder of this paper is structured as follows. Section 2 briefly overviews the related works. Section 3 presents a detailed description of our proposed gaze-dependent image re-ranking method for CBIR. The experimental results are presented in Section 4, where we provide qualitative and quantitative results of the proposed method. Section 5 discusses the implications of our findings and limitations associated with our study. Finally, we conclude with a summary of our contributions in Section 6.

2. Related Works

2.1. Semantic Gap in CBIR

The semantic gap is a significant challenge for CBIR models. It arises from the disparity between low-level visual features extracted from images and high-level semantic conceptions commonly utilized to depict the content of images. In other words, in CBIR, as images are the input, the ambiguity arising from the rich and complex content of a query image makes it challenging for the model to comprehend the inherent query intention embedded within the query image. For example, to retrieve similar images, discrepancies could appear between the similarity that the user envisioned and the similarity that the model perceived.

Recent research breakthroughs in deep learning [26–29] have opened up the possibility of solving the semantic gap problem in CBIR. One solution to address the above issue involves extracting subpatches from various regions of the query image and characterizing them using the in-depth features proposed by Razavian et al. [30]. By compressing the in-depth features of subpatches, more representative patch-based similarity can be computed to bridge the semantic gap. Wang et al. [31] proposed a two-stage approach for CBIR that combines sparse representation and feature fusion techniques to enhance the retrieval accuracy. In [32], the authors proposed a novel technique for CBIR by aggregating deep local features and generating compact global descriptors. Gong et al. [33] employed CNNs to extract deep characteristics from patches at different scales and areas, along with orderless pooling strategies. CNN-based approaches are typically superior to classic SIFT-based or GIST-based approaches [34,35] because they are preferred for feature extraction in image retrieval due to their ability to capture features closely related to the semantic attributes of images. Although CNNs are effective at extracting features for CBIR, their use typically involves analyzing the entire content of an image, which can result in inaccuracies when representing and identifying the underlying query intention within the query image.

Despite significant progress in recent years, the semantic gap remains a challenging problem for CBIR. An ongoing study is required to improve retrieval performance and develop more effective techniques for connecting low-level visual features with high-level semantic concepts.

2.2. Image Caption with Visual Information

To improve retrieval performance, many studies [24,36–38] have explored the introduction of various auxiliary information to generate image captions that are more in line with human habits and preferences, and visual information is one of them. He et al. [24] were dedicated to exploring the relevance between human attention and the discourse used in perception and text formation processes. They examined the mechanisms of attention allocation in the top-down soft attention method and proved the effectiveness of visual saliency for image captioning. In [38], a gazing sub-network was presented to estimate the gazing sequence from entities in a caption annotated by humans and then adopt the pointer network that automatically produces a similar description sequence imitating human visual habits.

Overall, the study of image captioning with visual information continues to evolve rapidly, and there is a growing interest in developing more sophisticated models that can generate captions that are not only accurate but also informative, diverse, and engaging.

2.3. Re-Ranking Method

In the domain of image retrieval, re-ranking is an essential component that entails the precise re-ordering of the initial retrieval results. Re-ranking is a crucial aspect of various retrieval tasks, including object retrieval, person re-identification, and CBIR, due to its ability to improve the accuracy and relevance of the retrieved results. Two distinct types of re-ranking approaches exist, namely self-re-ranking and interactive re-ranking.

Self-re-ranking is an automated technique that re-orders the initial retrieval results using data derived from the top-ranked images in the retrieval results. Certain retrieval approaches [39–41] employ text labels associated with top-ranked images as supplementary data to calculate the correlation between the query image and candidate images, thereby facilitating the re-ranking process. In contrast, Mandal et al. [42] utilized the class information of the top-ranked images during the re-ranking stage. The above techniques can enhance retrieval efficacy without requiring input from users. However, in the absence of supplemental information to bridge the discrepancy in comprehension between users and models, these methods still require assistance in disambiguating inherently vague queries. In contrast to the above techniques, our method bridges the semantic gap by incorporating gaze information to anticipate the user's preference.

Interactive re-ranking is a technique that involves re-ordering the initial retrieval results by interacting with users, typically through feedback mechanisms such as relevance feedback or users' preferences. Traditional interactive re-ranking techniques [43–47] entail users selecting images associated with the desired image from the initial retrieval results, after which the retrieved results are re-ordered based on the feedback obtained from users. In the fashion industry, novel approaches [48,49] have been proposed as alternatives for obtaining feedback, employing natural language to assess the top-ranked images. These approaches deviate from the conventional feedback mechanisms and provide a fresh outlook on evaluating fashion items. Chen et al. [50] propose a method that utilizes text feedbacks to improve image retrieval accuracy by incorporating multi-grained uncertainty regularization to handle the complex relationship between the image and text features. Re-ranking can be performed using reinforcement learning or deep metric learning, depending on the feedback received in a natural language format. Tan et al. [51] introduced an additional technique for efficiently handling multiple iterations of user-provided natural language queries to reorganize the retrieved results. Generally, interactive re-ranking uses users' diverse feedback to obtain re-ranking information. Furthermore, to better explore user preferences and bridge the semantic gap in CBIR, what type of information is suitable to be introduced as feedback while interacting with users is a topic worthy of in-depth exploration and long-term research.

3. Gaze-Dependent Image Re-Ranking

The proposed method comprises a set of sequential procedures: content-based image retrieval, gaze-dependent image captioning, and re-ranking. Figure 2 provides an overview of our method. First, we rank the candidate images using the conventional CBIR method based on the query image. Then, a gaze-dependent image caption is generated using gaze traces corresponding to the images obtained in the initial retrieval as the user's preference information. A relevant supplement should also be considered in case of accuracy degradation due to missing information about the query image. Finally, cross-modal image retrieval (CMIR) is performed with the generated captions to obtain images that better fit the user's preference as the re-ranking part.

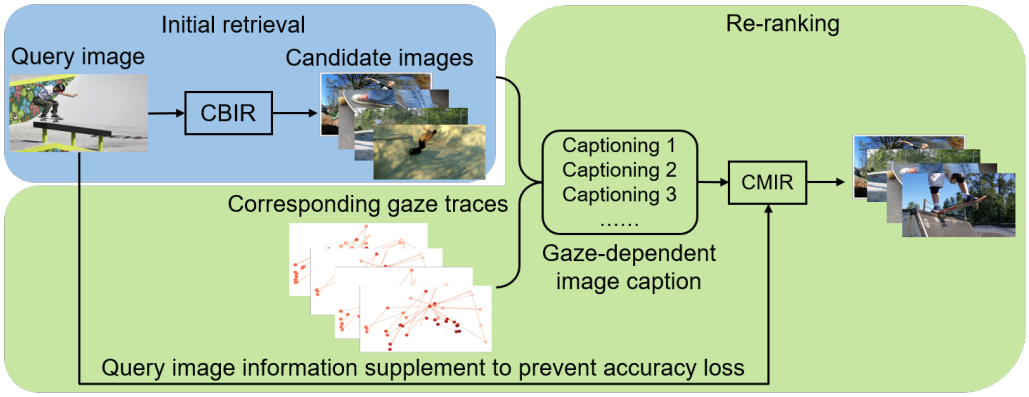


Figure 2. An overview of our method. We use gaze traces associated with initial retrieval results as feedback to generate gaze-dependent image captions. Moreover, we comprehensively use feature information from gaze-dependent image captions and query images for re-ranking.

3.1. Initial Retrieval

The proposed method takes the query image I_q as the input for the conventional CBIR model and ranks candidate images I_n (n = 1, ..., N; N represents the total number of candidate images) for the first step. As the proposed method necessitates the use of textual information to process captioning information during the re-ranking part, we expect that the CBIR model can handle both image and text elements naturally. Therefore, the proposed method focuses on cross-modal retrieval methods to perform the initial retrieval. Specifically, we introduce the CLIP model [23], which has been widely employed in contemporary image retrieval research and has exhibited significant retrieval precision.

The CLIP model employs a transformer-based encoder to create a shared latent space that facilitates the comparison of feature vectors obtained from both the query image and the candidate images. Furthermore, the initial results can be obtained by calculating their similarity $s_{q,n}$ as shown in the following equation.

$$s_{q,n} = \frac{\boldsymbol{v}_q \cdot \boldsymbol{v}_n}{|\boldsymbol{v}_q||\boldsymbol{v}_n|}, \tag{1}$$

where $\boldsymbol{v}_q$ and $\boldsymbol{v}_n$ represent the feature vectors of the query image and candidate images, respectively. The operator "·" represents the dot product that takes two vectors and returns a scalar value. The dot product of two vectors is defined as the product of their magnitudes and the cosine of the angle between them.

3.2. Gaze-Dependent Image Caption

To model image, captioning, and gaze information, we have constructed the CGT model, employing previous methods [22] to acquire knowledge of their interrelations. This model comprises three modules: an image encoder, a caption encoder–decoder, and a gaze trace encoder–decoder. As shown in Figure 3, each module is built on a transformer and receives $\mathbf{c}_v$, $\mathbf{c}_w$, $\mathbf{c}_r$ as input, denoting the features of image, caption, and gaze trace, respectively. The image encoder $\mathcal{T}_v$ is defined as follows:

$$\mathcal{T}_v = \mathrm{FFN}(\mathrm{MHA}_v(Q, V, K)), \tag{2}$$

where FFN(·) refers to the feed-forward network with two ReLU-based linear transformation layers, based on a previous study [52]. It also employs multi-head attention (MHA) with a weight matrix of query Q, value V, and key K to deal with the multi-modal input during training and is defined as follows:

$$\mathrm{MHA}_v(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_c)P^o, \tag{3}$$

$$\text{head}_i = \text{Attention}(QP_i^Q, KP_i^K, VP_i^V), \tag{4}$$

where P^o, P_i^Q, P_i^K, and P_i^V, the parameter matrices, are the projections following the definition [52]. As illustrated in Figure 3, it is worth noting that the outputs of the image encoder $\mathcal{T}_\text{v}$ have a significant impact on both the caption encoder–decoder $\mathcal{T}_\text{w}$ and the gaze trace encoder–decoder $\mathcal{T}_\text{r}$.

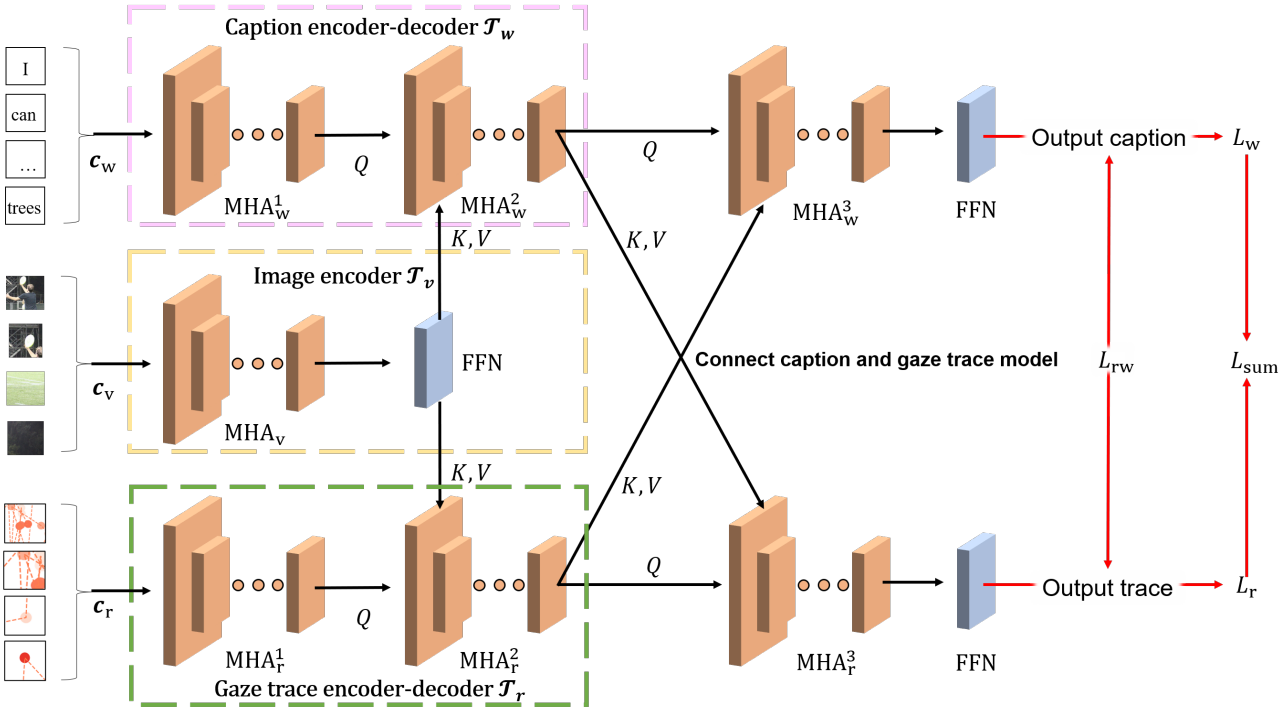


Figure 3. An overview of our connected caption and gaze trace (CGT) model. The model includes three transformer-based modules to learn the relevance of image, captioning, and gaze information. Note that the outputs of the image module $\mathcal{T}_\text{v}$ affect both the caption module $\mathcal{T}_\text{w}$ and the gaze trace module $\mathcal{T}_\text{r}$ to deliver the image features.

The CGT model is designed to effectively capture the interdependence between the image, caption, and gaze trace. Our method employs a symmetrical structure to generate captions and gaze traces to achieve this. The caption encoder–decoder $\mathcal{T}_\text{w}$ and the trace encoder–decoder $\mathcal{T}_\text{r}$ are defined as follows:

$$\mathcal{T}_\text{w} = \text{MHA}_\text{w}^2(\text{MHA}_\text{w}^1(\mathbf{c}_\text{w}, \mathbf{c}_\text{w}, \mathbf{c}_\text{w}), \mathcal{T}_\text{v}, \mathcal{T}_\text{v}), \tag{5}$$

$$\mathcal{T}_\text{r} = \text{MHA}_\text{r}^2(\text{MHA}_\text{r}^1(\mathbf{c}_\text{r}, \mathbf{c}_\text{r}, \mathbf{c}_\text{r}), \mathcal{T}_\text{v}, \mathcal{T}_\text{v}). \tag{6}$$

To train the CGT model, the total function L_all is defined as follows:

$$\mathcal{L}_\text{all} = \lambda_1 \mathcal{L}_\text{r} + \lambda_2 \mathcal{L}_\text{w} + \lambda_3 \mathcal{L}_\text{sum} + \lambda_4 \mathcal{L}_\text{rw}, \tag{7}$$

In this total function, $\mathcal{L}_\text{r}$ evaluates the difference between the bounding boxes of the predicted and actual gaze traces, which affects the performance of gaze trace generation. Conversely, $\mathcal{L}_\text{w}$ is a cross-entropy loss function that measures the dissimilarity between the human-generated and predicted captions. The total loss function, L_sum, is computed as the sum of $\mathcal{L}_\text{r}$ and $\mathcal{L}_\text{w}$. Furthermore, the model utilizes $\mathcal{L}_\text{rw}$, which is a cycle consistency loss that compares the output caption and trace. The weighting coefficients for these loss functions are represented as $\lambda*$, and their values were selected based on the method proposed in a previous study [53].

To summarize, using transformer-based architecture in our gaze-dependent image caption model offers several benefits. Since it is region-based, we can facilitate the extraction of even the slightest movement of the human gaze by dividing the image into smaller regions. Furthermore, under the training of the interdependent relationship between gaze information and visual/captioning features, our method can generate comprehensive captions controlled by the gaze trace.

3.3. Re-Ranking Based on Vision Information

After the above processing, the proposed method converts coordinate-format gaze trace data into text-format semantic information using image captions, which is used for a re-ranking that matches user preferences. To select the feedback information for re-ranking, obtaining many feedbacks from users contributes to reflecting their preferences. Conversely, too much feedback increases the burden for the user and the computational complexity of the model from a practical perspective. Considering the above factors, in this experiment, the proposed method uses the gaze traces corresponding to a specific number of images (the top a images) from the initial retrieval results as feedback from the user. The re-ranking process is as follows. First, the top a images from the initial retrieval results and their corresponding gaze traces are input into the CGT model to generate the same number of gaze-dependent image captions. By inputting the generated captions into the CLIP model, we can obtain the feature vectors v_1, v_2, v_3, ..., v_a representing user preference information. Moreover, we consider the query image feature during the re-ranking process to prevent accuracy degradation caused by the absence of its information. Finally, v_r used for re-ranking can be obtained by averaging the feature vector v_q of the query image with a feature vectors (v_1, v_2, v_3, ..., v_a) extracted from the gaze-dependent image captions by the following equation:

$$v_\mathrm{r} = \frac{v_\mathrm{q} + \sum_{k=1}^{a} v_k}{a+1}. \tag{8}$$

Because the major components that constitute the feature vector v_r

represent user preference information based on the gaze information, using v_r for re-ranking makes it possible to obtain the result ranking those images that reflect the user's preferences with a higher rank.

4. Experiment

4.1. Conditions

Dataset . In the experiments, we used pairs of images and human visual information from a dataset presented by [24] called HAIC. The dataset consists of 4000 images randomly chosen from the MSCOCO [25] and Pascal-50S [54] datasets. Each image has been annotated with a human-generated caption and corresponding gaze trace, indicating the regions of the image that the annotators focused on. With the training on this dataset, the CGT model can accurately learn the deep relationship between gazes, images, and captions to connect image features and the gaze trace using captioning to explicitly express semantic information, such as object relationships.

In our experiments, we divided the images in the HAIC dataset into training and test sets containing 2000 and 2000 images, respectively. Specifically, the 2000 images in the test set are randomly divided into 1500 candidate images and 500 query images in the test phase. Because each image in the HAIC dataset has a corresponding gaze trace and human-generated caption, we use the paired data of the same image separately in re-ranking during the test phase. Gaze traces of the candidate images are used as feedback information for re-ranking. In contrast, the human-generated caption of the query images is used as the ground truth to retrieve the desired image and perform the evaluation by calculating the metrics, both the initial retrieval results and the re-ranking.

Considering the balance between the introduction of sufficient gaze information and the calculation time of the network, we set a = 3 in this experiment, using the gaze trace of the top three images from the initial retrieval results to generate the gaze-dependent image

caption as the feedback information for re-ranking. For the $\lambda*$ in loss function, their values depended on the particular experiment and dataset being used. In this experiment, we set $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$, and $\lambda_4 = 0.1$, respectively.

To assess the effectiveness of our method, we applied recent CMIR and re-ranking approaches in the evaluation. Specifically, we compared our method against other existing techniques, such as [45,46,55,56], to assess its adaptability. The comparative methods were implemented using open-source codes provided by their respective authors. As for the time cost during training and retrieval, it is about 100 h and 70 s, respectively. Thus, it is necessary to consider the reduction of the computation cost for the retrieval phase.

4.2. Evaluation Metrics

To accurately evaluate the proposed method, we adopt two extensively used evaluation metrics, namely Recall@k and NDCG@k. Specifically, we set k to 1, 5, 10, 50, and 100 to measure the performance at different retrieval depths. Higher values of Recall@k and NDCG@k indicate the superior performance of the method. Recall@k is an extensively used evaluation metric for machine learning and information retrieval. It quantifies the capacity of a model to recognize all pertinent instances in a given dataset. The definition of Recall@k is shown as follows:

$$\text{Recall@}k = \frac{n_{\text{q}}}{k}, \tag{9}$$

where n_{q} represents the number of desired images that appeared in the top k retrieval results. The desired images refer to the retrieval results obtained using the human-generated caption corresponding to the query images in our experiment. NDCG@k (normalized discounted cumulative gain) is also an evaluation metric commonly used in image retrieval to measure the effectiveness of a ranked list of items. It measures the quality of the recommendations by assigning a score between 0 and 1 based on how well the recommended items match the user's preferences. NDCG considers each recommended item's relevance and position in the list of recommendations [57]. The relevance score of each item is usually determined by the user's feedback (e.g., ratings, clicks, purchases) on the items. NDCG is calculated by taking the discounted cumulative gain (DCG) of the ranked list and normalizing it by the ideal DCG(IDCG). Specifically, NDCG@k is defined as follows:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \tag{10}$$

where DCG is calculated by summing the relevance scores of the ranked items, discounted by their position in the list. IDCG is the maximum possible DCG for the list, calculated by assuming that the items are ranked in order of decreasing relevance. Specifically, DCG and IDCG are defined as follows:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \tag{11}$$

$$\text{IDCG@}k = \sum_{i=1}^{\min(k,|R|)} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \tag{12}$$

where rel_i denotes the relevance score of ith item in the desired image list.

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

To establish the adaptability of the proposed method, we conducted a comparative analysis with current state-of-the-art techniques such as those presented in [55,56]. The evaluation was performed using the HAIC datasets for both initial retrieval and re-ranking, and we present the corresponding results. Specifically, Tables 1 and 2 list the assessment results for Recall@k and NDCG@k, respectively. In the evaluation process, we first calculated the initial results on the basis of a CMIR model (e.g., CLIP), and then the proposed method was

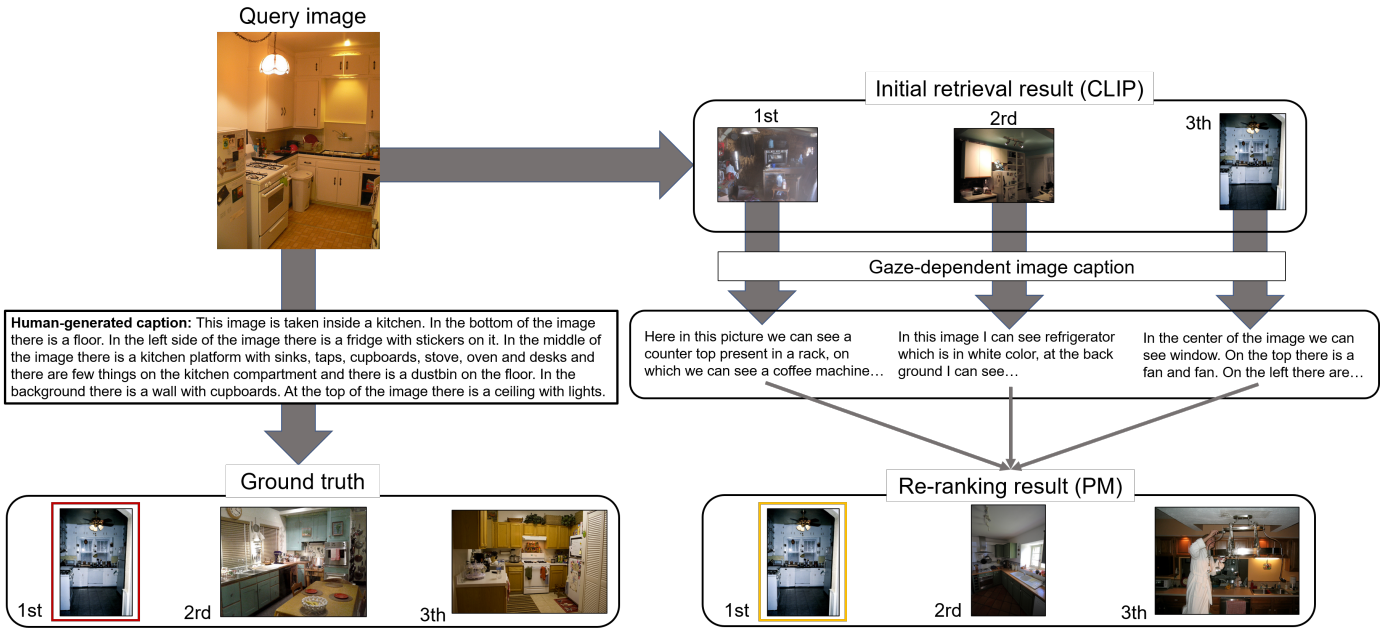
employed to rearrange the initial retrieval results. The results in Tables 1 and 2 illustrate that after re-ordering using the proposed method, all re-ranking results outperform the initial retrieval results obtained from their respective baseline (BL) methods. This means that the proposed re-ranking method can improve the performance of the initial retrieval results calculated by the BL method in CMIR. These findings demonstrate the efficacy of the proposed re-ranking method in enhancing CMIR performance. Examples of retrieval results are depicted in Figure 4a,b. To better evaluate the retrieval results qualitatively, instead of using the whole desired image list as the ground truth in the quantitative evaluation, in this study, we set the first image of the desired image list as the relevant image and observed the rank of this image in the re-ranking results. In Figure 4a, the proposed method re-ordered an image from the third position in the initial retrieval to the first position, the same rank as the ground truth of the same image. Furthermore, in Figure 4b, we find that the ground truth image was ranked low in the initial retrieval (not in the top 3) but re-ordered in a higher rank after being processed by the proposed method. Finally, we demonstrate some results of the generated gaze-dependent image captions and the original images overlaid with the corresponding gaze trace in Figure 5. As shown in Figure 5, it can be seen that the gaze controls the description order of the generated caption. For example, in Figure 5a, the gaze goes through the sky, forest, and grass, and the caption also follows this order to describe the image. Based on this qualitative assessment, it is verified that the proposed method can enhance the overall retrieval performance beyond its initial state. Furthermore, the efficacy of the re-ranking technique that incorporates eye gaze data is substantiated.

Table 1. Comparison of Recall@*k* between the proposed method and state-of-the-art methods. By comparing the results of applying the proposed method to multiple advanced cross-modal image retrieval methods, the general applicability of the proposed method to improve retrieval performance can be analyzed. The bold number in the table is the best result.

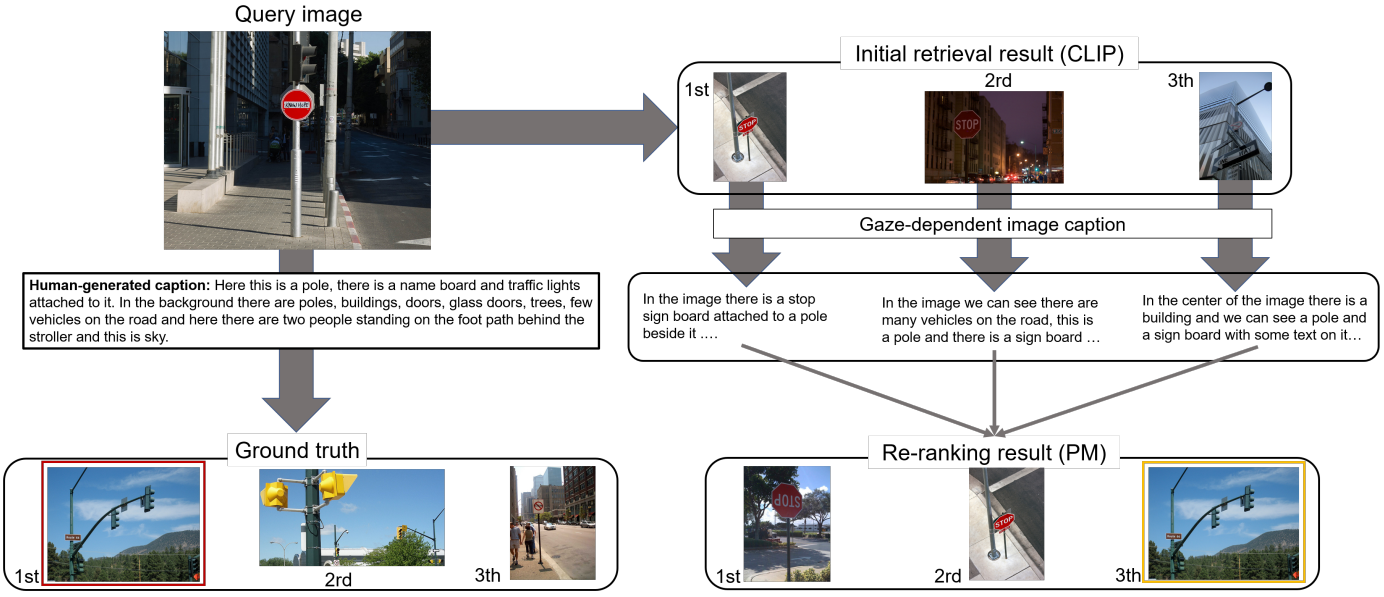
	Rcall@1	**Rcall@5**	**Rcall@10**	**Rcall@50**	**Rcall@100**
PVSE [55]	0.027	0.074	0.118	0.266	0.336
PM (PVSE)	**0.070**	**0.139**	**0.197**	**0.353**	**0.414**
Ji '19 [56]	0.026	0.074	0.119	0.273	0.341
PM (Ji '19)	**0.060**	**0.142**	**0.201**	**0.354**	**0.419**
CLIP [23]	0.080	0.220	0.298	0.423	0.479
PM (CLIP)	**0.120**	**0.304**	**0.360**	**0.493**	**0.575**

Table 2. Comparison of NDCG@*k* between the proposed method and state-of-the-art methods. The bold numbers in the table are the best results.

	NDCG@1	**NDCG@5**	**NDCG@10**	**NDCG@50**	**NDCG@100**
PVSE [55]	0.027	0.076	0.123	0.269	0.316
PM (PVSE)	**0.070**	**0.143**	**0.207**	**0.341**	**0.372**
Ji '19 [56]	0.026	0.078	0.125	0.273	0.317
PM (Ji '19)	**0.060**	**0.148**	**0.216**	**0.349**	**0.382**
CLIP [23]	0.080	0.232	0.328	0.489	0.545
PM (CLIP)	**0.120**	**0.317**	**0.398**	**0.549**	**0.629**



(**a**)



(**b**)

Figure 4. Retrieval results achieved using the proposed method. To demonstrate the qualitative evaluation, we set the first image in the ground truth as the relevant image and marked it with a red box. (**a**,**b**) are the re-ranking results for "kitchen" and "traffic lights" respectively. Correspondingly, we observed the rank of these images in the re-ranking results and highlighted them with a yellow frame.

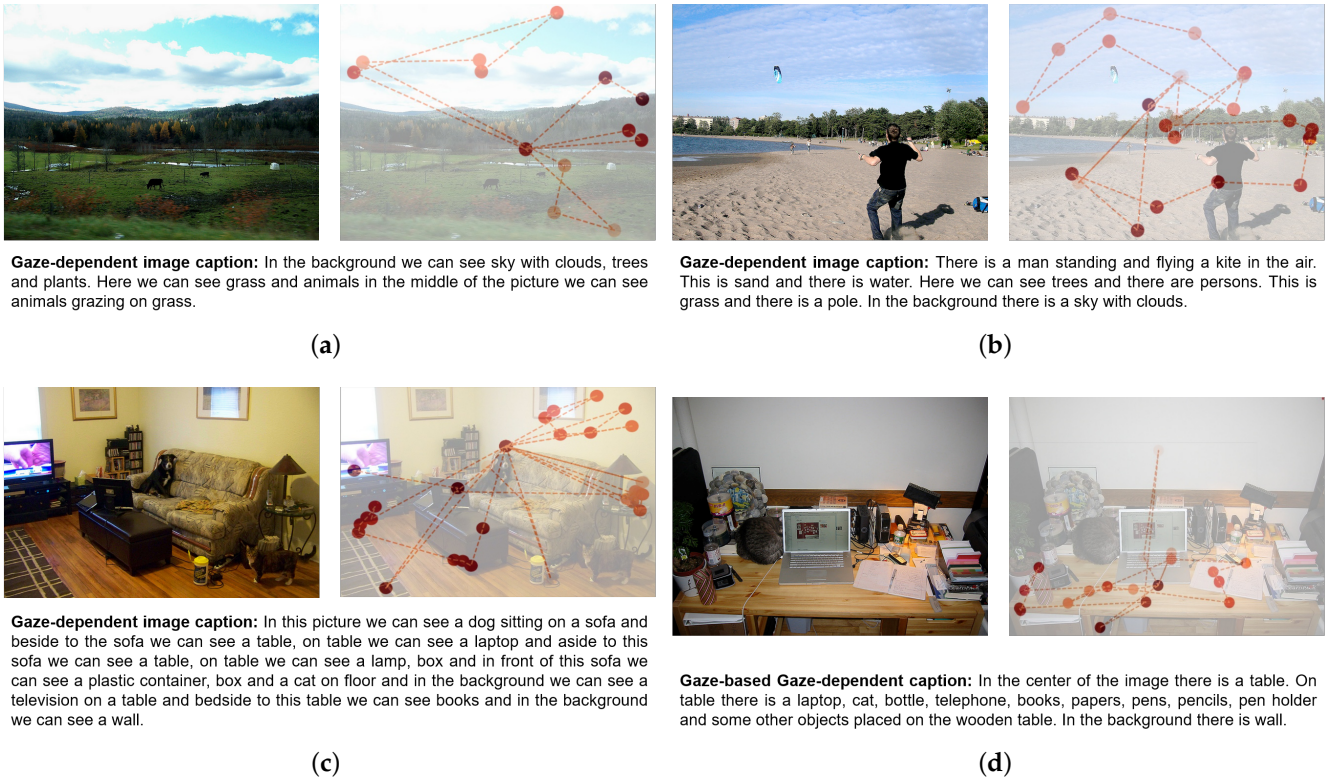


Figure 5. Examples of the original image and its corresponding gaze trace. In (**a**–**d**), the color of the dots in the gaze trace, from light to dark, represents the fixation order. The generated gaze-dependent caption is shown under the image.

4.4. Comparison with Re-Ranking Methods

In this subcategory, we perform an empirical analysis to assess the effectiveness of the proposed method compared with the conventional re-ranking method. Note that the proposed method uses a unique interactive re-ranking method, which converts coordinate-based gaze trace data into semantic information represented in a textual format using image captions. This method distinguishes itself from the conventional interactive re-ranking methods. Considering the above situation, it can be challenging to directly compare our proposed and conventional methods, as reported in [58]. To compare our proposed and existing methods, we evaluate our method against methods presented in [10,45,46], which can be assessed under the same conditions. Conventional techniques that obtain relevant feedback from users concerning the highest-ranked images are employed for the evaluation. Initially, we computed the initial retrieval based on the conventional CMIR model, CLIP [23]. Subsequently, we re-ordered the retrieval results using both the proposed method and the conventional re-ranking methods. Finally, we quantitatively compared the re-ranking results. Note that we utilized the hyper-parameters of the conventional interactive re-ranking methods that maximize the average retrieval efficacy.

Tables 3 and 4 illustrate the experimental results. In the BL column of Tables 3 and 4, the retrieval efficacy of the conventional CMIR approaches is presented. These tables show that the proposed method demonstrates superior performance over the conventional re-ranking methods in terms of retrieval effectiveness. Note that the conventional methods cannot significantly improve the BL retrieval performance. These results are likely due to the failure to determine the inherent query intention inside the query images. In contrast to the conventional re-ranking methods, our method uses gaze information feedback to facilitate re-ranking that accurately reflects user preferences. Based on these results, we could confirm the effectiveness of our method.

Table 3. Experimental results of Recall@*k* for interactive re-ranking methods. By comparing the evaluation metrics of the proposed method and multiple re-ranking methods, it is possible to quantify the improvement of the retrieval performance between the proposed method and other methods for the same task. The bold number in the table is the best result.

	Rcall@1	Rcall@5	Rcall@10	Rcall@50	Rcall@100
BL (CLIP [23])	0.080	0.220	0.298	0.423	0.479
Lin '15 [45]	0.086	0.251	0.316	0.447	0.526
putzu '20 [46]	0.109	0.284	0.349	0.463	0.551
polley '22 [10]	0.118	0.296	0.359	0.485	0.571
PM	**0.120**	**0.304**	**0.360**	**0.493**	**0.575**

Table 4. Experimental results of NDCG@*k* for interactive re-ranking methods. The bold numbers in the table are the best results.

	NDCG@1	NDCG@5	NDCG@10	NDCG@50	NDCG@100
BL (CLIP [23])	0.080	0.232	0.328	0.489	0.545
Lin '15 [45]	0.086	0.266	0.359	0.511	0.587
putzu '20 [46]	0.109	0.284	0.372	0.525	0.603
polley '22 [10]	0.118	0.303	0.386	0.542	0.609
PM	**0.120**	**0.317**	**0.398**	**0.549**	**0.629**

4.5. Ablation Study

In this subsection, we examine the influence of the transformation of the gaze trace data. The above experiments prove the versatility of our method; however, a compelling explanation of its novelty is still required, that is, introducing the gaze trace as the feedback of re-ranking and transforming it from coordinate data into a semantic format through image captioning. In such a situation, we require a comparative method that uses only the raw gaze data in the coordinate form for re-ranking. However, there are few relevant studies thus far, and it is not easy to make an effective comparison. Therefore, as depicted in Figure 6, the ablation method directly segments the sight-stayed areas of images according to the corresponding gaze data of the most focused regions (approximately five regions) and calculates the average feature vectors of these slices for re-ranking.

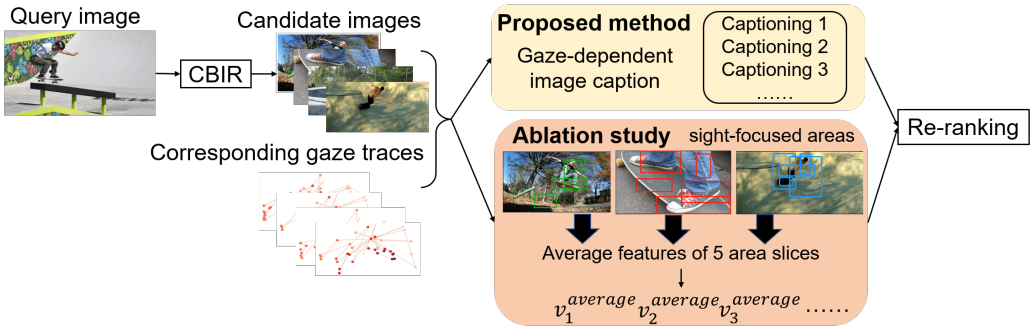


Figure 6. Concept diagram of the ablation study.

Tables 5 and 6 illustrate the experimental results. According to the results in these tables, our method outperforms the ablation methods. It is practical to generate gaze-dependent image captions to realize the re-ranking that accurately reflects user preferences.

Table 5. Experimental results of Recall@*k* for the verification of the novelty of our method to transform gaze trace from coordinate data into semantic information. Note that "Ablation" denotes the re-ranking directly based on the sight-stayed areas of images according to the corresponding gaze data without the image caption. The bold number in the table is the best result.

	Rcall@1	**Rcall@5**	**Rcall@10**	**Rcall@50**	**Rcall@100**
Ablation	0.090	0.216	0.285	0.452	0.534
PM	**0.120**	**0.304**	**0.360**	**0.493**	**0.575**

Table 6. Experimental results of NDCG@*k* for the verification of the novelty of our method. The bold numbers in the table are the best results.

	NDCG@1	**NDCG@5**	**NDCG@10**	**NDCG@50**	**NDCG@100**
Ablation	0.090	0.249	0.367	0.493	0.583
PM	**0.120**	**0.317**	**0.398**	**0.549**	**0.629**

5. Discussion

The proposed gaze-dependent re-ranking model enhances the ability to overcome the semantic gap. In this section, we will address the constraints of the current model and potential future directions for further research.

5.1. Limitation

The images that match the user's preferences will be ranked higher in the re-ranking results using the proposed method. However, there is still a problem in that the image at the top of the re-ranking result list is not always the desired image. Furthermore, in this study, the experiments were solely conducted under the condition where $a = 3$, and the efficacy of the proposed method under a broader range of environmental settings was not investigated. Lastly, as the proposed method attempts to establish the relationship between images, gazes, and captions via image captioning, the dataset utilized in the experiments must comprise data on the three models above. In this case, the current public dataset of gaze–image pairs required for adaptation is small. The lack of training data also leaves room for further improvement in the performance of the model. Some indications of this lack can be found in the results shown in Figure 5. We noticed that although the generated caption describes the content of the user's attention in the image in detail, there will also be problems of low text quality, such as repeated use of the same sentence pattern (as in Figure 5c) or simply stacking nouns (as in Figure 5d) to cause redundancy, which may affect the re-ranking performance.

5.2. Future Works

In future research, we intend to enhance the resilience of the proposed method, elevating the desired positions of images in the re-ranking results and ensuring the accuracy of the top-ranking results. Moreover, it is necessary to find the proper value a to explore the maximum likelihood balance between the adequate introduction of gaze information and the efficiency of the model during further experiments. Conversely, we need to find ways to expand the dataset, such as predicting the gaze trace for egovision images or videos with human annotations or referring to the data augmentation method [59] to increase existing saliency datasets. With sufficient training data, better performance in responding to complex and changing environments is expected. Furthermore, while the primary focus of this paper is to introduce a novel framework for retrieval using gaze-dependent captioning, the evaluation of the generated captions is an essential component of this application. As the quality of the captions directly impacts the retrieval performance of the model, it is imperative to conduct quantitative evaluations of the captioning in future research to explore ways to enhance the model's performance further. In practical applications, there may also be situations where a user performs retrieval repeatedly, so the multi-time

re-ranking problem for the proposed method will be our future study content. Finally, it is important to consider cases where images may be distorted or have low contrast. In such cases, it may be beneficial to use image restoration techniques as a pre-processing step to prioritize image clarity and obtain accurate gaze information in the future.

6. Conclusions

In this article, we have proposed a novel re-ranking method for CBIR to solve the CBIR semantic gap problem. We introduced gaze trace information as the user's feedback to predict user preference during image retrieval. Moreover, we transform the gaze data into semantic information using an image captioning method to help the model better comprehend the inherent query intention inside the query image. Specifically, the proposed method first generates gaze-dependent image captions on the basis of the user's gaze information corresponding to the images obtained in the initial retrieval as the user's preference information. Next, image retrieval is performed again using the generated captioning to obtain retrieval results that better reflect the user's preferences. The experimental results show that the proposed method effectively improves image retrieval accuracy by re-searching images using gaze trace information.

Author Contributions: Conceptualization, Y.F., K.M. and T.O.; methodology, Y.F.; data curation, Y.F.; writing—original draft, Y.F.; writing—review and editing, K.M. and T.O.; supervision, T.O. and M.H.; funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partly supported by JSPS KAKENHI Grant Number JP21H03456, JP20K19856, and AMED project Grant Number JP22zf0127004h0002.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available dataset was analyzed in this study. This data can be found here: https://github.com/SenHe/Human-Attention-in-Image-Captioning (accessed on 11 May 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wei, S.; Liao, L.; Li, J.; Zheng, Q.; Yang, F.; Zhao, Y. Saliency inside: Learning attentive CNNs for content-based image retrieval. *IEEE Trans. Image Process.* **2019**, *28*, 4580–4593. [CrossRef]
2. Latif, A.; Rasheed, A.; Sajid, U.; Ahmed, J.; Ali, N.; Ratyal, N.I.; Zafar, B.; Dar, S.H.; Sajid, M.; Khalil, T. Content-based image retrieval and feature extraction: A comprehensive review. *Math. Probl. Eng.* **2019**, *2019*, 9658350. [CrossRef]
3. Dubey, S.R. A decade survey of content based image retrieval using deep learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2687–2704. [CrossRef]
4. Alsmadi, M.K. Content-based image retrieval using color, shape and texture descriptors and features. *Arab. J. Sci. Eng.* **2020**, *45*, 3317–3330. [CrossRef]
5. Garg, M.; Dhiman, G. A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants. *Neural Comput. Appl.* **2021**, *33*, 1311–1328. [CrossRef]
6. Shen, Y.; Qin, J.; Chen, J.; Yu, M.; Liu, L.; Zhu, F.; Shen, F.; Shao, L. Auto-encoding twin-bottleneck hashing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2818–2827.
7. Wang, R.; Wang, R.; Qiao, S.; Shan, S.; Chen, X. Deep position-aware hashing for semantic continuous image retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Village, CO, USA, 1–5 March 2020; pp. 2493–2502.
8. Li, X.; Yang, J.; Ma, J. Recent developments of content-based image retrieval (CBIR). *Neurocomputing* **2021**, *452*, 675–689. [CrossRef]
9. Enser, P.; Sandom, C. Towards a comprehensive survey of the semantic gap in visual image retrieval. In Proceedings of the Image and Video Retrieval: Second International Conference 2003, Urbana-Champaign, IL, USA, 24–25 July 2003; Proceedings 2; Springer: Berlin/Heidelberg, Germany, 2003; pp. 291–299.
10. Polley, S.; Mondal, S.; Mannam, V.S.; Kumar, K.; Patra, S.; Nürnberger, A. X-Vision: Explainable Image Retrieval by Re-Ranking in Semantic Space. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, 17–22 October 2022; pp. 4955–4959.
11. Fadaei, S. New dominant color descriptor features based on weighting of more informative pixels using suitable masks for content-based image retrieval. *Int. J. Eng.* **2022**, *35*, 1457–1467. [CrossRef]

12. Zhang, X.; Jiang, M.; Zheng, Z.; Tan, X.; Ding, E.; Yang, Y. Understanding image retrieval re-ranking: A graph neural network perspective. *arXiv* **2020**, arXiv:2012.07620.
13. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1318–1327.
14. Zhang, D.; Guo, G.; Wang, H.; Fujian, C. Image Retrieval Method Based on Two Models Re-Ranking (IRM2R). 2022 .Available online: https://scholar.google.co.jp/scholar?q=Image+Retrieval+Method+Based+on+Two+Models+Re-Ranking&hl=zh-CN&as_sdt=0&as_vis=1&oi=scholart (accessed on 22 March 2023).
15. Xu, Y.; Gao, S.; Wu, J.; Li, N.; Yu, J. Personalized saliency and its prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2975–2989. [CrossRef] [PubMed]
16. Moroto, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. Few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention. *Sensors* **2020**, *20*, 2170. [CrossRef]
17. Moroto, Y.; Maeda, K.; Ogawa, T.; Haseyama, M. User-centric visual attention estimation based on relationship between image and eye gaze data. In Proceedings of the 2018 IEEE 7th Global Conference on Consumer Electronics, Nara, Japan, 9–12 October 2018; pp. 73–74.
18. Sugano, Y.; Ozaki, Y.; Kasai, H.; Ogaki, K.; Sato, Y. Image preference estimation with a data-driven approach: A comparative study between gaze and image features. *J. Eye Mov. Res.* **2014**, *7* . [CrossRef]
19. Bernardi, R.; Cakici, R.; Elliott, D.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N.; Keller, F.; Muscat, A.; Plank, B. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.* **2016**, *55*, 409–442. [CrossRef]
20. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.
21. Tuli, S.; Dasgupta, I.; Grant, E.; Griffiths, T.L. Are Convolutional Neural Networks or Transformers more like human vision? *arXiv* **2021**, arXiv:2105.07197.
22. Yuhu, F.; Keisuke, M.; Takahiro, O.; Miki, H. Human-Centric Image Retrieval with Gaze-Based Image Captioning. In Proceedings of the 2022 IEEE International Conference on Image Processing, Bordeaux, France, 16–19 October 2022; pp. 3828–3832.
23. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, Switzerland, 18–24 July 2021 ; PMLR: New York, NY, USA, 2021; pp. 8748–8763.
24. He, S.; Tavakoli, H.R.; Borji, A.; Pugeault, N. Human attention in image captioning: Dataset and analysis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8529–8538.
25. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 806–813.
31. Wang, W.; Jiao, P.; Liu, H.; Ma, X.; Shang, Z. Two-stage content based image retrieval using sparse representation and feature fusion. *Multimed. Tools Appl.* **2022**, *81*, 16621–16644. [CrossRef]
32. Babenko, A.; Lempitsky, V. Aggregating deep convolutional features for image retrieval. *arXiv* **2015**, arXiv:1510.07493.
33. Gong, Y.; Wang, L.; Guo, R.; Lazebnik, S. Multi-scale orderless pooling of deep convolutional activation features. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part VII 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 392–407.
34. Zhang, S.; Yang, M.; Cour, T.; Yu, K.; Metaxas, D.N. Query specific rank fusion for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 803–815. [CrossRef]
35. Zhou, D.; Li, X.; Zhang, Y.J. A novel CNN-based match kernel for image retrieval. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.
36. Murrugarra-Llerena, N.; Kovashka, A. Cross-modality personalization for retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 6429–6438.
37. Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; Weston, J. Engaging image captioning via personality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 12516–12526.
38. Alahmadi, R.; Hahn, J. Improve Image Captioning by Estimating the Gazing Patterns from the Caption. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 1025–1034.

39. Wang, T.; Xu, X.; Yang, Y.; Hanjalic, A.; Shen, H.T.; Song, J. Matching images and text with multi-modal tensor fusion and re-ranking. In Proceedings of the 27th ACM international conference on multimedia, Nice, France, 21–25 October 2019; pp. 12–20.
40. Wei, W.; Jiang, M.; Zhang, X.; Liu, H.; Tian, C. Boosting cross-modal retrieval With MVSE++ and reciprocal neighbors. *IEEE Access* **2020**, *8*, 84642–84651. [CrossRef]
41. Yu, X.; Chen, T.; Yang, Y.; Mugo, M.; Wang, Z. Cross-modal person search: A coarse-to-fine framework using bi-directional text-image matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
42. Mandal, D.; Biswas, S. Query specific re-ranking for improved cross-modal retrieval. *Pattern Recognit. Lett.* **2017**, *98*, 110–116. [CrossRef]
43. Giacinto, G. A nearest-neighbor approach to relevance feedback in content based image retrieval. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 456–463.
44. Liang, S.; Sun, Z. Sketch retrieval and relevance feedback with biased SVM classification. *Pattern Recognit. Lett.* **2008**, *29*, 1733–1741. [CrossRef]
45. Lin, W.C.; Chen, Z.Y.; Ke, S.W.; Tsai, C.F.; Lin, W.Y. The effect of low-level image features on pseudo relevance feedback. *Neurocomputing* **2015**, *166*, 26–37. [CrossRef]
46. Putzu, L.; Piras, L.; Giacinto, G. Convolutional neural networks for relevance feedback in content based image retrieval: A Content based image retrieval system that exploits convolutional neural networks both for feature extraction and for relevance feedback. *Multimed. Tools Appl.* **2020**, *79*, 26995–27021. [CrossRef]
47. Xu, B.; Bu, J.; Chen, C.; Wang, C.; Cai, D.; He, X. EMR: A scalable graph-based ranking model for content-based image retrieval. *IEEE Trans. Knowl. Data Eng.* **2013**, *27*, 102–114.
48. Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; Feris, R. Dialog-based interactive image retrieval. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 678–688. [CrossRef]
49. Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.J.; Fei-Fei, L.; Hays, J. Composing text and image for image retrieval-an empirical odyssey. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 6439–6448.
50. Chen, Y.; Zheng, Z.; Ji, W.; Qu, L.; Chua, T.S. Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization. *arXiv* **2022**, arXiv:2211.07394.
51. Tan, F.; Cascante-Bonilla, P.; Guo, X.; Wu, H.; Feng, S.; Ordonez, V. Drill-down: Interactive retrieval of complex scenes using natural language queries. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 2647–2657.
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
53. Meng, Z.; Yu, L.; Zhang, N.; Berg, T.L.; Damavandi, B.; Singh, V.; Bearman, A. Connecting what to say with where to look by modeling human attention traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12679–12688.
54. Vedantam, R.; Zitnick, C.L.; Parikh, D. Collecting image description datasets using crowdsourcing. *arXiv* **2014**, arXiv:1411.3041.
55. Song, Y.; Soleymani, M. Polysemous visual-semantic embedding for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 1979–1988.
56. Ji, Z.; Wang, H.; Han, J.; Pang, Y. Saliency-guided attention network for image-sentence matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5754–5763.
57. Wang, Y.; Wang, L.; Li, Y.; He, D.; Liu, T.Y. A theoretical analysis of NDCG type ranking measures. In Proceedings of the Conference on Learning Theory, Princeton, NJ, USA, 12–14 June 2013; PMLR: New York, NY, USA, 2013; pp. 25–54.
58. Rossetto, L.; Gasser, R.; Lokoč, J.; Bailer, W.; Schoeffmann, K.; Muenzer, B.; Souček, T.; Nguyen, P.A.; Bolettieri, P.; Leibetseder, A.; et al. Interactive video retrieval in the age of deep learning–detailed evaluation of VBS 2019. *IEEE Trans. Multimed.* **2020**, *23*, 243–256. [CrossRef]
59. Che, Z.; Borji, A.; Zhai, G.; Min, X.; Guo, G.; Le Callet, P. How is gaze influenced by image transformations? Dataset and model. *IEEE Trans. Image Process.* **2019**, *29*, 2287–2300. [CrossRef] [PubMed]