Anti-Confounding Hashing: Enhancing Radiological Image Retrieval via Debiased Weighting and Counterfactual Reasoning

Yajie Zhang, Yao Hu, *Member, IEEE*, Chengjun Cai, *Member, IEEE*, Yu-An Huang, *Member, IEEE*, Zhi-An Huang, *Member, IEEE*, and Kay Chen Tan, *Fellow, IEEE*

***Abstract*— Content-based medical image retrieval (CBMIR) enables physicians to make evidence-based diagnoses by retrieving similar medical images and recalling previous cases stored in databases. However, existing CBMIR models are prone to capturing superficial correlations due to confounding factors such as complex host organs and lesions, imaging discrepancies, artifacts, and inconsistent protocols. To address this issue, we propose a plug-and-play anti-confounding hashing (ACH) method, which uses debiased sample weighting and lesion counterfactual reasoning (LCR) to directly capture the natural direct effect (NDE) of lesions on query medical images without bias. The devised debiased weighting (DBW) loss adopts a backdoor adjustment to separate lesions from confounders. To effectively locate salient areas of lesions, we present a coarse-to-fine lesion positioning (C2F-LP) module by counterfactual reasoning. On two real-world radiological image datasets, ACH achieves 0.2%–9% improvement in mean average precision (mAP) over the six state-of-the-art methods, when using code lengths ranging from 8-bit to 32-bit. Its robustness to confounding factors is demonstrated through explainable visual analysis.**

***Index Terms*— Confounding factors, content-based medical image retrieval (CBMIR), counterfactual reasoning, COVID-19, debiased weighting (DBW), hashing.**

I. Introduction

RADIOLOGICAL imaging is committed to providing high-quality, noninvasive scans for various purposes, including screening, diagnosis, assessment of treatment response, and monitoring disease recurrence [1]. In the realm of medical image retrieval, querying extensive databases empowers physicians to access a wealth of case information, thereby enhancing the precision in identifying lesions, anomalies, and diseases. Simultaneously, medical image retrieval serves as a robust data resource for medical research, facilitating a deeper understanding of disease progression and treatment outcomes. Therefore, the integration of medical image retrieval not only addresses the challenges posed by the scarcity of experienced radiologists and established standards but also stands as a facilitative tool in radiological imaging.

With the tremendous advancements in GPU technology and image processing, content-based medical image retrieval (CBMIR) has attracted substantial attention over the past decade. Of the numerous existing CBMIR methods, hashing-like functions are widely utilized, thanks to their high storage efficiency and low computational complexity. These methods transform each sample into a compact binary code, allowing efficient similarity searches through XOR and bit-count operations in the Hamming space. Hence, the utilization of hash functions in medical image retrieval technology exhibits promising potential for positively impacting efficiency enhancement and reduction in storage costs.

Medical image retrieval presents greater difficulties than natural image retrieval due to the intricate tissue textures and anatomical structures present in medical images, making it challenging to identify small abnormal regions [2]. Hashing-based CBMIR can be broadly classified into two groups: unsupervised and supervised methods. Recent developments in contrastive learning [3], [4] and autoencoder [5] have shown promising results in learning invariant hash codes without labeled data. Supervised methods, which incorporate label dependency, have technical advantages in semantic correlation measures to generate similarity-preserving hash codes, offering greater potential for downstream applications. For example, lesion detection methods [6] can integrate with existing object detection algorithms, segmentation modules, and attention mechanisms [7], [8] to precisely target pathological areas. Multiview and multiscaled feature learning methods [9] have been proposed to learn complementary information for feature fusion from different perspectives. Despite their strong performance in CBMIR, these methods face challenges in medical data acquisition. Traditional methods require time-consuming manual intervention and the growing

Received 14 February 2023; revised 8 February 2024 and 17 September 2024; accepted 27 December 2024. Date of publication 13 January 2025; date of current version 6 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62202399, Grant U21A20512, and Grant 6220239; in part by the Research Grants Council of the Hong Kong, SAR, under Grant PolyU11211521, Grant PolyU15218622, Grant PolyU15215623, and Grant C5052-23G; in part by the Fundamental Research Funds for the Central Universities under Grant G2023KY05102; in part by the General Program of National Natural Science Foundation of China under Grant 62472353; and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515011984 and Grant 2023A151514013. *(Corresponding authors: Yu-An Huang; Zhi-An Huang.)*

Yajie Zhang is with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, SAR, and also with the Research Office, City University of Hong Kong (Dongguan), Dongguan 523000, China (e-mail: yajie.zhang@connect.polyu.hk).

Yao Hu and Kay Chen Tan are with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, SAR (e-mail: yaohu4-c@my.cityu.edu.hk; kctan@polyu.edu.hk).

Chengjun Cai and Zhi-An Huang are with the Research Office, City University of Hong Kong (Dongguan), Dongguan 523000, China (e-mail: chengjun.cai@cityu-dg.edu.cn; huang.za@cityu-dg.edu.cn).

Yu-An Huang is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710000, China (e-mail: yuanhuang@nwpu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2025.3526760

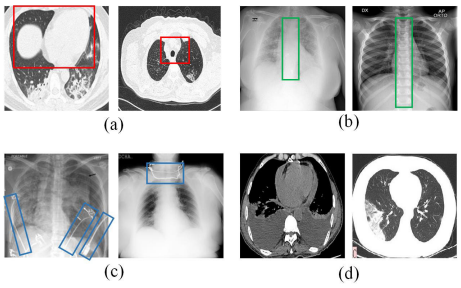

Fig. 1. Illustration of confounding factors present in radiological image retrieval. (a) and (d) CT images of COVID and (b) and (c) X-ray images of COVID. (a) Co-occurrence of complex host organs and lesions, (b) imaging discrepancies by different devices, (c) conspicuous imaging artifacts, and (d) inconsistent acquisition protocols.

volume of data strains management systems. Additionally, the issue of confounding bias arising from the data acquisition process needs to be addressed in large-scale radiological image datasets.

The visualization in Fig. 1 highlights four common confounding factors presented in X-ray and CT images of coronavirus disease 2019 (COVID-19).

1) The presence of complex host organs and lesions [shown in red boxes, Fig. 1(a)] can vary depending on the scanning position, causing significant differences in the anatomical manifestation of images within the same class.
2) The imaging discrepancies caused by different devices can lead to a mismatch in spectrum and style, causing a domain shift. As shown in Fig. 1(b), the varying penetrating power of X-rays can also affect the clarity of the host spine (in green boxes).
3) The conspicuous imaging artifacts [marked in blue boxes, Fig. 1(c)] can interfere with the pathological areas and compromise the generation of hashing codes.
4) The inconsistent acquisition protocols pose a challenge to the reproducibility and generalizability of feature modeling. This can be observed in the global variations in color and contrast as seen in Fig. 1(d) where images are scanned from different windows, such as the pulmonary window (left) and the mediastinal window (right).

The emergence of confounding factors leads to the dissimilarity of samples in the same class due to different confounding factors, while samples from different classes exhibit similarity due to shared confounding factors. In a nutshell, eliminating the influence of uncontrolled confounding factors is crucial for hash learning to explore the causal relationship between data and labels, paving the way for the achievement of precise retrieval techniques.

To solve the above issues, this study proposes the anti-confounding hashing (ACH) method to enhance radiological image retrieval. The proposed method utilizes natural direct effect (NDE) [10] to generate hash codes that accurately reflect the contrast of lesions while removing the impact of confounder exposure on the outcome. This allows NDE to measure the expected change in hash code when a lesion goes from absence to presence. To generate counterfactual samples accurately, a debiased weighting (DBW) loss based on backdoor adjustment is presented to remove the spurious correlations between lesions and confounding factors. A lesion counterfactual reasoning (LCR) module is also designed to generate the counterfactual hash codes by a two-stage coarse-to-fine lesion positioning (C2F-LP) process. This process first coarsely identifies the lesion area in the image and then refines the mask through a fine-tuning process. The hashing network is trained using both the original and counterfactual samples, forcing it to focus on real causes and thus improving retrieval performance. By introducing the ACH method, it promises several significant benefits to society, especially in the healthcare sector. First, as medical imaging enters the era of big data, the ACH method can help in managing and navigating large datasets, making it easier to extract meaningful insights without the distortions caused by confounding factors. Second, the ACH method can alleviate the burden on radiologists by reducing the dependency on manual annotation and by providing a more efficient way to retrieve relevant images. Third, the plug-and-play nature of ACH allows for seamless integration with existing hashing-based data management systems. ACH is designed to be scalable, allowing it to handle large volumes of data and adapt to changing data acquisition needs. This can facilitate a deeper understanding of disease progression and treatment outcomes.

The main contributions of ACH can be summarized as follows.

1) A novel ACH method is proposed to explore the NDE of lesions on hash code generation, marking the first attempt to address confounding effects in medical image retrieval (as per the knowledge of the authors).
2) Based on backdoor adjustment, DBW is designed to optimize the hashing network, allowing it to learn anti-confounding hash codes. The weight of confounders with larger deviations is assigned more significance, ensuring that the network prioritizes them in the optimization process.
3) The proposed LCR module is designed to generate masks for counterfactual samples in the LCR module, resulting in precise lesion detection.
4) The proposed ACH method is versatile and can be easily incorporated into existing hashing methods. Extensive evaluations, conducted using six baseline models, demonstrate the effectiveness of ACH in terms of various evaluation metrics.

The rest of this article is organized as follows. Section II is a review of related work in the field. Section III outlines the pipeline of baseline hashing methods. The proposed ACH method is introduced in Section IV. The experimental results are presented in Section V. Finally, the conclusion of this work is presented in Section VI.

II. Related Work

This section provides the context for this work by reviewing the principles, categories, and development of hashing as the basis for CBMIR. Afterward, we delve deeper into the current state of hashing algorithms for medical image retrieval. Finally, we examine previous methods for addressing confounding biases in medical image analysis.

A. Content-Based Image Hashing

The goal of a hashing algorithm is to maintain similarity and structural consistency between the original data space and the Hamming space. To achieve this, hashing algorithms focus on feature mapping and similarity measurement.

Based on the method of feature mapping, hashing algorithms can be divided into traditional and deep hashing methods. Since traditional hashing methods map handcrafted features to hash codes, they have difficulty capturing semantic information from complex images. For example, locality-sensitive hashing [11] encodes the handcraft features to hash codes by random projections and permutations. Iterative quantization [12] uses the principal component analysis (PCA) as the mapping function and minimizes the quantization loss by mapping data to the vertices of a zero-centered binary hypercube. Cross-modal info-max hashing [13] aims to maximize the mutual information between the binary code and input features. Binary set embedding [14] maps samples into a common Hamming space in a way that each sample is represented by a set of local feature descriptors from cross-domains.

Deep hashing methods, on the other hand, leverage neural network-based mapping functions to capture high-level semantic information of large-scale datasets, which can be categorized into metric learning-based approaches and feature augmentation-based approaches. Metric learning-based methods focus on simultaneously learning a metric space during hash code learning, intending to minimize the distances between similar samples and maximize the distances between dissimilar samples in this space. These methods commonly utilize metric learning loss functions (LFs) such as pairwise loss [15], [16], [17], triplet loss [18], and center loss [19], [20], [21]. On the other hand, feature augmentation-based methods prioritize enhancing hash code expressiveness by introducing additional information or conducting feature augmentation on the original features. For instance, joint learning of global and local features [22] or the utilization of hierarchical features [23] can be employed in hash learning. The inherent hashing learning capabilities of these methods position them as foundational techniques for ACH. However, their specificity regarding medical images is limited, overlooking the intricacies and diversity present in medical imaging.

B. Medical Image Retrieval Methods

CBMIR primarily focuses on identifying pathology-bearing regions (PBRs) through three main aspects: multitask learning, multiview learning, and graph representation learning. The objective of multitask learning is to concurrently learn multiple related tasks, extracting a shared feature representation across tasks to enhance the model's generalization capability and performance. Y-Net [24] generates compact hash codes from visual features in PBRs by combining segmentation loss and classification loss. Multiview learning aims to enhance the understanding and modeling capabilities by integrating information from multiple perspectives. LAGE-Net [6] uses hierarchical agglomerative clustering to aggregate patches of a histopathological whole slide image into subregions and identifies lesion areas using a self-attention mechanism. DGMFE [25] leverages the complementary information from functional and anatomical features to encode heterogeneous data into unified hash codes through a multimodality graph. Graph representation learning employs graph networks to establish relationships between different regions, thereby constructing associations among various regions. For instance, Zheng et al. [26] utilizes the pathologist's browsing path as prior knowledge to construct a graph of regions of interest for histopathology image retrieval.

C. Early Efforts to Combat Confounding Bias

Recent efforts have been made to reduce the effects of confounding bias in medical images [27]. For example, C-CAM [28] utilizes backdoor adjustment [10] to reduce the confounding bias caused by the co-occurrence of different host organs. TraCE [29] generates counterfactual samples by adding lesions to normal images through an encoder–decoder structure. Lenis et al. [30] implemented counterfactual reasoning by inpainting approach applied on salient areas. However, these approaches do not fully control for the effect of confounding factors among samples, leading to an over- or under-estimate of the association between observed confounders and health conditions. To address this issue, this work proposes to tackle confounding bias in CBMIR from two perspectives: DBW and counterfactual reasoning. The proposed ACH approach differs significantly from the aforementioned methods in several aspects: 1) unlike TraCE, which is limited to binary classification tasks, ACH can also be applied to multiclass causal inference, thereby broadening its scope of application; 2) ACH relies solely on label-level weak supervision, significantly eliminating the necessity for labor-intensive manual intervention traditionally required for training a hash model with causal recognition capabilities; and 3) ACH employs DBW loss at an early stage to disentangle lesions from confounding factors. This is a crucial step to help in effectively mitigating the adverse impacts of confounding factors on counterfactual sample generation, where other compared approaches are less effective.

III. Preliminaries: Baseline Pipeline

To clarify the methodology of the proposed method, the supervised hashing pipeline is depicted in Fig. 2(middle). This pipeline mainly consists of three stages. In the first stage, the hashing network converts medical images into feature embeddings and subsequently generates compact binary hash codes from these embeddings. The second stage focuses on developing a suitable LF that effectively guides the learning process of the hashing network. The third stage involves employing appropriate optimization techniques, such as the discrete optimization method for ADSH [16] or stochastic gradient descent optimization for CSQ [19], to refine the generated hash codes and enhance the overall performance of the hashing model. Table I details the specific optimization method for each baseline. Throughout the article, matrices are represented using bold uppercase letters (e.g., $\mathbf{A}$) and vectors are represented using bold lowercase letters (e.g., $\mathbf{a}$).

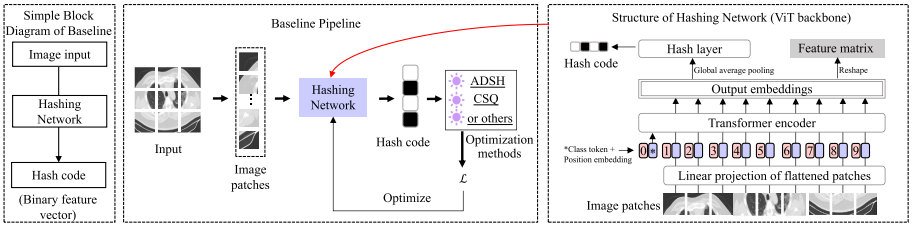

Fig. 2. Pipeline of the baseline hashing model. The diagram illustrates the process of converting an image into a compact hash code representation using the hashing network, followed by optimization with utilized LFs.

The ith row vector of a matrix $\mathbf{A}$ is denoted as $\mathbf{a}_i$, and the jth value of a vector $\mathbf{a}_i$ is represented as $\mathbf{a}_{ij}$.

A. Problem Formulation

Suppose that a dataset $D_s = \{(x_i, \mathbf{y}_i)|i = 1, \ldots, n\}$ is given, where n is the number of samples, x_i is the ith image, and $\mathbf{y}_i \in \mathbb{R}^C$ is the corresponding label. y_{ic} is a binary value that represents whether x_i belongs to the cth category. The goal of the hashing method is to learn a hash function $h : x \to \mathbf{b} \in \{-1, 1\}^q$ that maps the original image x into a compact q-bit hash code.

B. Hashing Network

The hashing network in the proposed method is made up of two components: a backbone for image embedding and a hash layer for hash code generation. The backbone is typically either a convolutional neural network (CNN) or a Vision Transformer. In this work, we choose the Vision Transformer model, as it is effective in capturing long-range associated features in medical images [31]. The chosen Vision Transformer model is the commonly used ViT model [32]. As shown in Fig. 2(right). The ViT model takes a set of 1-D token embeddings as input and follows the standard transformer architecture. To adapt to the image data, the input image $x \in \mathbb{R}^{(H\cdot P)\times(W\cdot P)\times L}$ is first reshaped into a sequence of flattened 2-D patches $x_p \in \mathbb{R}^{N\times(P^2\cdot L)}$, where H and W are the numbers of patches in the height and width direction of the image, L represents the number of channels of the original image, (P, P) denotes the resolution of each patch, and the total number of square patches is $N = H \times W$. Then, the image patches are transformed into patch embedding $\mathbf{G}$ of D dimensions through a trainable linear projection as

$$\begin{aligned}\mathbf{G} &= f_{\text{lp}}\left(x_p; x_{\text{class}}, \mathbf{E}_{\text{lp}}, \mathbf{E}_{\text{pos}}\right) \\ &= \left[x_{\text{class}}; x_p^1\mathbf{E}_{\text{lp}}; x_p^2\mathbf{E}_{\text{lp}}; \cdots; x_p^N\mathbf{E}_{\text{lp}}\right] + \mathbf{E}_{\text{pos}} \end{aligned} \tag{1}$$

where f_{lp} is the function of linear projection, $x_{\text{class}} \in \mathbb{R}^D$ is the class embedding that aggregates the contents of image patches, $\mathbf{E}_{\text{lp}} \in \mathbb{R}^{(P^2\cdot L)\times D}$ is a learnable matrix for feature transformation, and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ represents the position embedding that reflects the path position to the image. Next, the patch embedding $\mathbf{G}$ is passed through the Transformer encoder structure with multiheaded self-attention (MSA), layer normalization (LN), and multilayer perceptron (MLP) modules. The Transformer encoder outputs the embedding vector of class and patch token $\mathbf{F}$ as

$$\mathbf{F} = f_b(\mathbf{G}; \theta_b) \in \mathbb{R}^{(N+1)\times D} \tag{2}$$

where θ_b denotes the parameters of the Transformer encoder f_b. Finally, $\mathbf{F}$ is subjected to a global average pooling (GAP) layer to avoid overfitting. The output of the GAP layer is then fed into a hash layer, which is implemented as a fully connected layer with a Tanh activation function (f_h). We can obtain the hash code $\mathbf{b}$ as follows:

$$\mathbf{b} = f_h\left(\text{GAP}\left(\mathbf{F}_{1:(N+1),:}\right); \theta_h\right) \in [-1, +1]^q \tag{3}$$

where θ_h denotes the parameter set of f_h, $\mathbf{F}_{0,:}$ is the class embedding, and $\mathbf{F}_{1:(N+1),:}$ represents the feature vectors of all image patches except for the class embedding.

C. Baseline LF

The goal of supervised hashing methods is to encode similar (visually or semantically) images into similar binary codes with minimized Hamming distance. This is achieved by utilizing label information and representation learning. To evaluate the generated hash codes, metric-based LFs are employed, of which the pairwise loss is a commonly used one [15], [16], [19], [33], [34], [35]. In this work, the pairwise loss is adopted as the baseline LF for illustration purposes, but it is worth noting that other types of metric LFs (e.g., center LF and class-balanced pairwise function) can also be used in our framework in a flexible, plug-and-play manner. The basic idea behind the pairwise LF is to minimize the Hamming distance of similar data pairs and maximize the Hamming distance of dissimilar data pairs. The calculation of the pairwise loss can be formulated as

$$\mathcal{L}_b = \sum_{i=1}^{n} \left\|\mathbf{b}_i^T\mathbf{V} - \hat{\mathbf{s}}_i\right\|^2 \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{(\sum_{c=1}^{C} n_c)\times q}$ represents the set of hash codes from all categories, with n_c being the number of samples in the cth class and $\hat{\mathbf{s}}_i$ is a vector of similarity labels between $\mathbf{b}_i$ and $\mathbf{V}$. In existing supervised hashing methods, the definition of the set of $\mathbf{V}$ can vary. For instance, in DSH, the hash codes of the current batch are defined as $\mathbf{V}$, while in ADSH, the hash codes of the database are defined as $\mathbf{V}$.

IV. Method

In this section, we present the causality-effect analysis in CBMIR and outline the process of transforming the estimation of NDE into two subproblems. The first subproblem ensures that the similarity of the generated hash codes closely aligns with the similarity labels. The second subproblem focuses on accurately generating counterfactual samples to provide

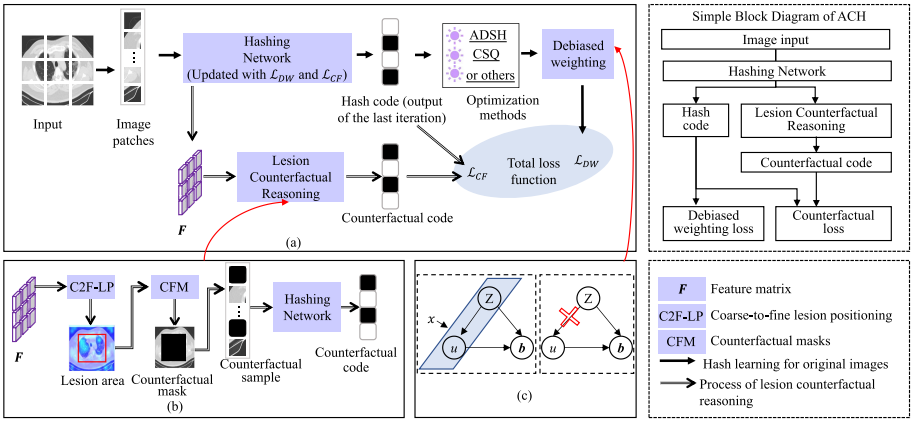

Fig. 3. Architecture of ACH is composed of three main components: a baseline pipeline, a DBW loss, and an LCR module. The baseline pipeline inputs an image into the hashing network to produce a hash code. The DBW loss balances the confounding factors across samples, while the LCR module generates counterfactual samples by adding counterfactual masks, as determined by C2F-LP, to the original images. (a) ACH pipeline. (b) LCR. (c) DBW.

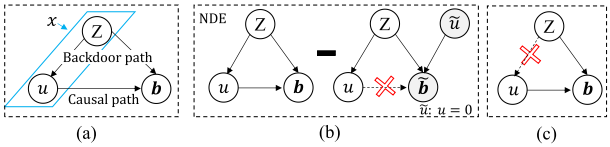

Fig. 4. (a) Causal graph for CBMIR consists of the input image x, the x's invariant feature u, and x's confounders Z, and the hash code $\mathbf{b}$ of x. The relationship between these variables is represented in the graph. (b) NDE is used to measure the expected changes when u transitions from presence to absence. In contrast to the natural indirect effect, which considers the transition of Z from presence to absence, the NDE is more directly associated with the causal pathway from u to the outcome. (c) Process of backdoor adjustment is illustrated to break the link between u and Z so that the relationship between the input image x, the invariant feature u, and the hash code $\mathbf{b}$ can be properly estimated.

a clearer understanding of causal relationships. The framework of ACH is illustrated as Fig. 3(a), which consists of a baseline pipeline, a DBW loss, and an LCR module. The DBW [Fig. 3(c)] integrated into the pipeline addresses the first subproblem by enhancing the learning of hash codes in alignment with similarity labels through backdoor adjustment. The LCR module [Fig. 3(b)] tackles the second subproblem by facilitating the generation of counterfactuals that reflect potential alterations in the data. We will introduce the theoretical foundations of the two subproblems in Section IV-A, provide details on the DBW in Section IV-B, and elaborate on the specifics of the LCR in Section IV-C.

A. Cause–Effect Analysis in CBMIR

The causal relationship between image x and its generated hash code $\mathbf{b}$ is depicted using a structural causal model in Fig. 4. This relationship is represented using a structural causal model. The content of image x is comprised of two elements: the invariant feature u and the confounders Z. The causal relationships between these elements are explained using the causal theory presented in [10].

As shown in Fig. 4(a), the presence of confounders has an impact on the representation of the invariant feature, which is represented by the arrow $Z \rightarrow u$. Additionally, there is a direct dependence of $u \rightarrow \mathbf{b}$ and an indirect dependence of $Z \rightarrow \mathbf{b}$. The confounders interfere with estimating the direct effect of the invariant feature on the hash code through the backdoor path $u \leftarrow Z \rightarrow \mathbf{b}$. To estimate the direct effect of the invariant feature on the hash code, the confounders need to be controlled, which is challenging in practice.

In this work, as shown in Fig. 4(b), the principle of NDE is utilized to estimate the direct effect of u on $\mathbf{b}$. The NDE measures the change in $\mathbf{b}$ as u changes from presence (represented as $u = 1$) to absence (represented as $u = 0$) while conditioning on Z. This allows the hashing network to estimate the direct effect of the invariant feature on the hash code while holding the confounders steady. The calculation of NDE can be formulated as

$$\text{NDE}(Z) = P(\mathbf{b}|u = 1, Z) - P(\mathbf{b}|u = 0, Z). \tag{5}$$

Substituting (5) into (4), we can obtain the optimization function $\mathcal{L}_b$ as

$$\begin{aligned}\mathcal{L}_b &= \sum_{i=1}^{n} \left\| (\mathbf{b}_i(u = 1, Z) - \mathbf{b}_i(u = 0, Z))^T \mathbf{V} - \hat{\mathbf{s}}_i \right\|^2 \\ &= \sum_{i=1}^{n} \left\| \left(\mathbf{b}_i^T(u = 1, Z)\mathbf{V} - \mathbf{s}_i\right) - \left(\mathbf{b}_i^T(u = 0, Z)\mathbf{V} - \tilde{s}_i\right) \right\|^2\end{aligned} \tag{6}$$

where $\hat{\mathbf{s}}_i$ is divided into the original similarity label $\mathbf{s}_i$ and counterfactual similarity label $\tilde{s}_i$. One solution to minimize $\|(a - b) - (c - d)\|^2$ is $a - b = c - d = 0$. Therefore, the optimization in (6) can be transformed into two subproblems

$$\mathcal{L}_1 = \sum_{i=1}^{n} \left\| \left(\mathbf{b}_i^T \mathbf{V} - \mathbf{s}_i\right) \right\|^2 \tag{7}$$

and

$$\mathcal{L}_2 = \sum_{i=1}^{n} \left\| \left(\tilde{b}_i^T \mathbf{V} - \tilde{s}_i\right) \right\|^2 \tag{8}$$

where we use $\mathbf{b}_i$ and $\tilde{b}_i$ to simplify $\mathbf{b}_i(u = 1, Z)$ and $\mathbf{b}_i(u = 0, Z)$, respectively. Equation (7) represents the general LF in traditional deep hashing algorithms. To accurately estimate NDE, it is crucial to generate $\tilde{b}_i$ and estimate $\tilde{s}_i$ with precision. This is reflected in (8).

The key to accurately generating $\tilde{b}_i$ lies in the accurate detection of u, usually represented by the lesion, in the input image x. However, the existence of the backdoor path $u \leftarrow Z \rightarrow \mathbf{b}$ entangles u and Z, making it difficult to accurately detect u. It is, therefore, necessary to separate the relationship between u and Z before estimating NDE. To achieve this, a DBW strategy based on backdoor adjustment is proposed in Section IV-B to break the link between u and Z during the generation of $\mathbf{b}$. Then, in Section IV-C, LCR is introduced to detect u using C2F-LP to generate $\tilde{b}$ and estimate $\tilde{s}$. In essence, ACH aims to uncover the relationship between u and $\mathbf{b}$ through DBW and LCR, by breaking apart the entanglement between u and Z and effectively managing Z.

B. Sample DBW

The traditional deep hashing methods are prone to confounding bias because of the existence of the backdoor path $u \leftarrow Z \rightarrow \mathbf{b}$, which leads to a pseudo-correlation between u and Z as shown in Fig. 4(a). To address this issue, the sample DBW is introduced to disconnect u and Z through backdoor adjustment. This method aims to ensure that all possible values of $Z = \{z\}$ are equally likely to occur with u, thereby breaking the link between u and Z. The backdoor adjustment on u can be formulated as

$$P(\mathbf{b}|\mathrm{do}(u)) = \sum_{z, z \in Z} P(\mathbf{b}|u, z)P(z) \tag{9}$$

where $\mathrm{do}(u)$ signifies the causal intervention on u and $P(z)$ is a constant for all confounders.

To address the confounding bias in the dataset, backdoor adjustment is employed in the sample DBW process. Fig. 5(a) describes the components of a sample in category c, which consists of an invariant feature u^c and a set of confounders $Z^c = \{z_i^c\}, 1 \leq i \leq m$. With m confounding factors affecting u^c, the hash codes of the cth class can be represented by the set $\mathbf{b}(u^c, z_1^c), \cdots, \mathbf{b}(u^c, z_m^c)$. However, it is important to note that the probabilities of these confounding factors are not equal, meaning that $P(z_1^c) \neq \cdots \neq P(z_m^c)$. As a result, $P(z_i^c)$ cannot act like a constant $P(z)$ in backdoor adjustment. To balance the probabilities of confounding factors, the DBW process assigns a higher weight to samples with less frequent confounders and a lower weight to samples with frequent confounders [as depicted in Fig. 5(b)].

The composition of samples belonging to the same category is identical in terms of their invariant feature u^c but different in terms of their confounders z_i^c. The differences between samples in the same category are mostly attributed to their differing confounders. Samples containing frequent confounders are more similar to other samples in their category compared to samples with infrequent confounders. As such, the frequency of a confounding factor can be used to calculate the intraclass average distance of samples that contain it. Given the ith sample in the cth category, the intraclass average distance is calculated as r_{ci}

$$r_{ci} = \frac{\sum_{j=1}^{n_c} \left\| \mathbf{b}_{ci}^T \mathbf{b}_j / q - 1 \right\|^2}{n_c} \tag{10}$$

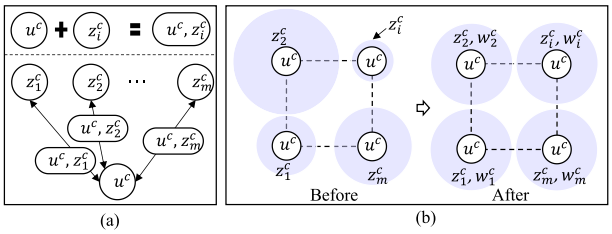

Fig. 5. (a) Composition of samples belonging to the cth category: each sample in the cth category is composed of an invariant feature u^c and a set of confounders $Z^c = \{z_i^c\}, 1 \leq i \leq m$. (b) DBW based on backdoor adjustment: the proportions of confounders can be visualized as the size of azure solid circles in which they are located. To balance the impact of various confounders, a weight w_i^c is assigned to each confounder z_i^c based on its proportion.

where q represents the length of the hash code. A larger value of r_{ci} indicates that the sample contains a confounder with a lower proportion, and thus deserves more attention from the hashing network. The weight of a sample is determined by applying the softmax operation on $\mathbf{r}_c$ as

$$\mathbf{w}_c = \mathrm{softmax}(\mathbf{r}_c). \tag{11}$$

This results in w_{ci} which can be added to the baseline loss in (7) to arrive at the final DBW loss L_{DW} as

$$\mathcal{L}_{\mathrm{DW}} = \sum_{c=1}^{C} \sum_{i=1}^{n_c} w_{ci} \left\| \mathbf{b}_{(ci)}^T \mathbf{V} - \mathbf{s}_{(ci)} \right\|^2 \tag{12}$$

where $\mathbf{s}_i$ is reached as

$$s_{ij} = \begin{cases} 1 & \text{if } \mathbf{y}_i{}^T \mathbf{y}_{\mathbf{v}_j} \geq 1 \\ -1 & \text{otherwise.} \end{cases} \tag{13}$$

C. Lesion Counterfactual Reasoning

We present the LCR method for detecting lesion areas in images. This method implements $u = 0$ for generating $\tilde{b}_i$ in (8) and estimating $\tilde{\mathbf{s}}_i$. The C2F-LP algorithm is introduced for precise lesion positioning through a two-step process: first, the lesion is roughly located using a sliding window on the activity map and then the exact lesion is identified using random key points on the target area. The counterfactual sample is generated by masking the detected lesion, and its Hamming distance from the original sample should be maximized to improve hash code generation, which is achieved by optimizing the counterfactual loss (8). The workflow of this process is illustrated in Fig. 6.

1) Coarse-to-Fine Lesion Positioning: The first stage of C2F-LP uses the sliding window method on the calculated active map to roughly locate the salient pathological area. The feature matrix $\mathbf{F}$ is transformed into feature maps $\mathbf{E}$ with dimensions $H \times W \times D$, where $N = H \times W$. To locate the salient area, the active map $\mathbf{A} \in \mathbb{R}^{H \times W}$ is calculated by averaging the values of the feature maps along the channel direction as

$$A_{hw} = \frac{\sum_{d=1}^{D} E_{dhw}}{D} \tag{14}$$

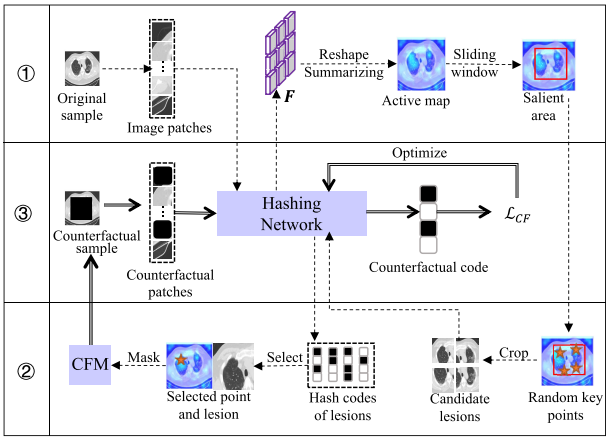

Fig. 6. Workflow of the LCR process consists of three steps. Steps ① and ② represent the two-stage C2F-LP. Step ①: The feature matrix F generated by the hashing network is used to preliminary locate the salient area by applying a sliding window method on the calculated active map. Step ②: The hashing network then evaluates potential candidate lesions surrounding the target salient area by selecting key points at random and choosing the most probable lesion and corresponding key point for the CFM process. Step ③: The processes in Steps ① and ② are repeated iteratively, allowing the hashing network to optimize its hash code generation by minimizing the counterfactual loss. This LF enforces dissimilarity between the original and counterfactual samples, allowing the hashing network to effectively distinguish between them.

where A_{hw} is the (h, w)th element of $\mathbf{A}$. Then, the sliding window M^{W} is employed to detect the salient area $\overline{\mathbf{A}}$ as

$$\overline{\mathbf{A}}_{(\tilde{X},\tilde{Y},\tilde{H},\tilde{W})} = \arg\max_{\overline{\mathbf{A}}} \left(\sum \left(\mathbf{A} \ominus \mathrm{M}^{\mathrm{W}} \right) \right) \tag{15}$$

where $(\mathbf{A} \ominus \mathrm{M}^{\mathrm{W}})$ refers to a cutting operation performed on $\mathbf{A}$ using M^{W} to identify the lesion position, and $\tilde{X}$, $\tilde{Y}$, $\tilde{H}$, and $\tilde{W}$ are the coordinates of the center point of $\overline{\mathbf{A}}$ and the predefined height and width of $\overline{\mathbf{A}}$, respectively.

In the second stage of the lesion detection process, the target pathological area is divided into a number of candidate lesion segments. To accomplish this, a set of K randomly located key points $\{P_1, P_2, \cdots, P_K\}$ are first selected in the salient area. The x- and y-coordinates of each key point, P_k, are defined as $\tilde{X}_k \in [\tilde{X} - (\tilde{W}/2), \tilde{X} + (\tilde{W}/2)]$ and $\tilde{Y}_k \in [\tilde{Y} - (\tilde{H}/2), \tilde{Y} + (\tilde{H}/2)]$, respectively. These key points are then used as center points to generate the candidate lesion segments $\{I_1, I_2, \cdots, I_K\}$, which have a height of $\tilde{H}^P \in [1, \tilde{H}]$ and width of $\tilde{W}^P \in [1, \tilde{W}]$. Each of these segments, I_k, can be represented with a hash code obtained from a hashing function, f, as follows:

$$\mathbf{b}_{I_k} = f(I_k; \theta). \tag{16}$$

The appropriate segment P_k is determined by evaluating the corresponding hash codes using intraclass variation as follows:

$$P_{(\tilde{X}_P,\tilde{Y}_P)} = P_{k(\tilde{X}_k,\tilde{Y}_k)} \quad \text{s.t.} \arg\min_{\mathbf{b}_{\mathrm{I_k}}} \frac{\sum_{j=1}^{n_c} \left\| \mathbf{b}_{I_k}{}^T \mathbf{b}_j / q - 1 \right\|^2}{n_c}. \tag{17}$$

This criterion is applied to ensure that the selected lesion segment has the closest Hamming distance of hash codes compared to other original samples that belong to the same class.

2) Counterfactual Reasoning: In the third stage, counterfactual masking (CFM) is applied to the original image by using multiscale masks centered at the selected key point. The aim of this is to generate multiple masked samples that have diverse representations. The multiscale masks are used to fully cover the pathological area, as a single fixed-size mask may not be sufficient. The counterfactual mask $\tilde{\mathbf{M}} \in \mathbb{R}^{1\times H\times W}$ is represented as

$$\tilde{M}_{1hw} = \begin{cases} 0 & \text{if } w \in \left[\tilde{Y}_P - \frac{s_w}{2}, \tilde{Y}_P + \frac{s_w}{2} \right], \\ & \text{and } h \in \left[\tilde{X}_P - \frac{s_h}{2}, \tilde{X}_P + \frac{s_h}{2} \right]; \\ 1 & \text{otherwise} \end{cases} \tag{18}$$

where (s_h, s_w) represents the height and width of the mask. To adapt to the input of ViT, this 3-D counterfactual mask is flattened to 1-D sequences with the size of $N(H \times W)$ and then is utilized to generate counterfactual sample $\tilde{x}_p$ on the image patches as

$$\tilde{x}_p = \mathrm{reshape}\left(\tilde{\mathbf{M}}\right) \otimes x_p \tag{19}$$

where $\otimes$ denotes the dot product operation. Finally, following (3), the counterfactual sample is encoded into a hash code by the hashing network as

$$\tilde{\mathbf{b}} = f\left(\tilde{x}_p; \theta\right). \tag{20}$$

To evaluate the synthetic weight of the counterfactual sample $\tilde{s}_{ij}$, we consider that after masking the lesion information, $\tilde{\mathbf{b}}_i$ should be as different from $\mathbf{v}_j$ that originally belongs to the same category as possible. With this consideration, $\tilde{s}_{ij}$ is defined as follows:

$$\tilde{s}_{ij} = \begin{cases} 0 & \text{if } \mathbf{y}_{\tilde{x}_i} = \mathbf{y}_{\mathbf{v}_j} \\ \dfrac{\sum_{t=1}^{n_c} \tilde{\mathbf{b}}_i{}^T \mathbf{v}_t}{n_c} & \text{otherwise} \end{cases} \tag{21}$$

where $\mathbf{y}_{\tilde{x}_i}$ represents the original label of $\tilde{x}_i$, and n_c is the number of samples in the cth category that the ith sample belongs to. To address the issue of deep hashing models being overwhelmed by frequent confounders, we give more weight to counterfactual hash codes that have frequent confounders when compared with the original sample. This helps the hash network to accurately distinguish between the frequent confounders and the lesions. The weight of a counterfactual sample is calculated as

$$\tilde{w}_{ci} = \tilde{\mathbf{r}}_c.\max() - \tilde{r}_{ci} \tag{22}$$

where $\tilde{\mathbf{r}}_c.\max()$ is the maximum value of $\tilde{\mathbf{r}}_c$. Likewise, the softmax operation is used to normalize $\tilde{\mathbf{w}}_c$. Finally, $\tilde{\mathbf{w}}_c$ can be integrated with (8) to reach the counterfactual reasoning loss $\mathcal{L}_{\mathrm{CF}}$

$$\mathcal{L}_{\mathrm{CF}} = \sum_{c=1}^{C} \sum_{i=1}^{n_c} \tilde{w}_{ci} \sum_{j=1}^{n} \left\| \tilde{\mathbf{b}}_{(ci)}{}^T \mathbf{V} - \tilde{\mathbf{s}}_{(ci)} \right\|^2. \tag{23}$$

As such, the total LF is written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{DW}} + \mathcal{L}_{\mathrm{CF}}. \tag{24}$$

Algorithm 1 Pseudo-Code of ACH

Input: Training dataset D_t, query dataset D_q, database D_{db}
Outpt: Produced hash codes $\mathbf{B}_q$ of D_q, $\mathbf{B}_{db}$ of D_{db}

Pretrain the backbone f_b with a self-supervised task (MAE [36]) and a classification task on D_t
Initialize hash network $f = f_h(f_b(\cdot))$
for i = 1, ..., Epoch **do**
  $X \leftarrow$ Select(D_t) // Select a batch of data from D_t
  $\mathcal{F} = \{\mathbf{F}\} \leftarrow f_b(X)$ // Compute output embeddings
  $\mathbf{B} \leftarrow f_h(\mathcal{F})$ // Compute hash codes
  $\mathcal{L}_1 \leftarrow (\mathbf{b}_i, \mathbf{V}, \mathbf{s}_i)$ via Eq. (7) // Calculated by baselines
  # Sample Debiased Weighting
  $\mathbf{w}_c \leftarrow \{\mathbf{b}_{cj} | j = 1, \cdots, n_c\}$ via Eq. (10) & Eq. (11)
  $\mathcal{L}_{DW} \leftarrow \mathrm{DBW}(\mathbf{w}_c, \mathcal{L}_1)$ via Eq .(12)
  # Lesion Counterfactual Reasoning
  $\tilde{M} = \{\tilde{M}\} \leftarrow$ C2F-LP($\mathcal{F}$, f) // Counterfactual masks
  $\tilde{X}_p \leftarrow (\tilde{M}, X)$ // Generate counterfactual samples
  Calculate $\tilde{B} = f(\tilde{X}_p)$ to derive $\tilde{S}$ by Eq. (21)
  Calculate $\mathcal{L}_{CF}$ by Eq. (23) // Counterfactual loss
  Update f by minimizing $\mathcal{L} = \mathcal{L}_{DW} + \mathcal{L}_{CF}$
end for
$\mathbf{B}_{db} \leftarrow f(D_{db})$
Query phase
$\mathbf{B}_q \leftarrow f(D_q)$
$\{.\} \leftarrow \arg\min Hamming(\mathbf{B}_q, \mathbf{B}_{db})$
Prioritize the retrieved candidate samples in $\{.\}$

D. Query Phase

After the training phase (offline), the query phase (online) just needs to compare the query image only with similar images based on the well-trained hashing network f. It is worth noting that the hashing network, trained using DBW and LCR, has already been endowed with the capability of causal discovery. In the online phase, we discard these two modules and solely utilize the backbone of the hashing network for generating query hash codes. This is advantageous in reducing the query time complexity and storage complexity during the online stage. We can obtain the hash code as follows:

$$\mathbf{b} = \mathrm{sign}(f(x; \theta)). \tag{25}$$

The priority in the retrieval of candidate samples is given to those with a minimal Hamming distance to the query image. To ease the understanding of ACH, the pseudo-code is presented as Algorithm 1.

V. Experiments

In this section, comprehensive experiments are conducted to evaluate ACH. First, the datasets, baselines, evaluation protocols, and implementation details are introduced. Second, the comparisons between baselines and ACH are presented across several widely used evaluation metrics. Third, we also conduct some exploration experiments to analyze the parameter sensitivity and the potential and scalable applications, that is, the ability of transfer learning (TL) and the retrieval task by using lesion areas. Finally, the visualization results of the C2F-LP module and salient maps are provided to help better understand the ACH.

TABLE I
Details of Six Baseline Methods. "Q" Represents the Query Set and "DB" Represents the Database

Method	LF	CB	HCO	LFE
DSH	Pairwise	False	Discrete	False
HashNet	Pairwise	True	Continuous	False
ADSH	Pairwise	True	Continuous (Q), Discrete(DB)	False
CSQ	Center	False	Continuous	False
DBDH	Pairwise	False	Discrete	False
TransHash	Pairwise	True	Continuous	True

A. Datasets

We evaluate the proposed method on three public radiological datasets: COVID-19 Radiograph [37], COVID-CT-Dataset [38], Brain Tumor [39], and a dataset of thermo-graphic IR images [40]. In the following, the details of each dataset are given.

COVID-19 Radiograph contains approximately 21 000 images in four categories, that is, COVID, lung opacity, normal, and viral pneumonia. We randomly select 4200 images as the query set, with the remaining images as the database. In the training process, we randomly choose 2000 images from the database as the training set.

COVID-CT-Dataset has 746 CT scan images in two categories (349 case samples and 397 control samples). We randomly divide the dataset into the query set with 149 images and the database with 594 images. In the training process, all samples in the database are used to train the hashing network.

Brain Tumor collects 3264 images for nuclear MRI in four categories, that is, glioma tumor, meningioma tumor, no tumor, and pituitary tumor. We follow the official train/test split to distribute 2879 and 394 images for the database set and query set, respectively.

Dataset of thermo-graphic IR images includes 369 images from 11 categories. The dataset is randomly partitioned into training and test sets in an 8:2 ratio.

The first three medical datasets have skewed class proportions. In this work, we utilize the COVID-19 Radiograph and COVID-CT-Dataset to comprehensively verify the performance of ACH. The dataset of thermo-graphic IR images is utilized to evaluate the adaptability of ACH beyond radiographic images. Thermal imaging infrared images are widely used in fields such as electrical system diagnosis and fault detection [41]. Afterward, the Brain Tumor dataset is used to evaluate the effectiveness of TL from X-ray or CT images to MRI images.

B. Baselines and the Evaluation Protocol

To assess the effectiveness of ACH, we compare it with six state-of-the-art deep hashing methods, which serve as benchmark methods. These methods include DSH (CVPR'16) [15], HashNet (ICCV'17) [33], ADSH (AAAI'18) [16], CSQ (CVPR'20) [19], DBDH (NeuroComputing'20) [34], and TransHash (ICMR'22) [35]. Table I summarizes the major differences among benchmark methods, viewed from four perspectives, that is, the adopted LF, hash code optimization (HCO), and whether consider class balance (CB) issue or

TABLE II
NETWORK AND IMPLEMENTATION DETAILS OF THE BASELINE METHODS AND ACH

Basic information					
Backbone		ViT-Base	Optimizer		AdamW
Image size		128*128	Patch size		8
Pre-training with a self-supervised method: MAE [36]					
Epoch	500	Batch size	64	Learning rate	1e-3
Pre-training with a classification task					
Epoch	150	Batch size	64	Learning rate	1e-3
Basic parameters for baseline methods and ACH					
Epoch		150	Learning rate		1e-5
Batch size		64	Weight decay		0.05
Hyper-parameters for ACH					
$\tilde{H}\&\tilde{W}$		12 & 12	$\tilde{H}^P\&\tilde{W}^P$		12 & 12
K	4	(s_h, s_w)		{(2,2), (4,4), (6,6), (8,8)}	

local feature extraction (LFE). The source codes of DSH, HashNet, ADSH, CSQ, and DBDH used were kindly provided by the authors. Additionally, we reproduced the source code of TransHash according to the original paper. The network and implementation details of baselines and ACH are listed in Table II for a fair comparison.

We evaluate the proposed method in terms of three widely used metrics, that is, mean average precision (mAP), top-K precision curve (P@K), and precision–recall (PR) curve. mAP reflects the mAP of the retrieved images over the database. Top-K precision indicates the proportion of the relevant images in the top-K retrieved images. The PR curve shows the tradeoff between precision and recall by varying the thresholds. The larger the metric value represents the better the performance of the method. In addition, the two-sided Wilcoxon sign-ranked test (a nonparametric test) is used to verify whether the performance improvement achieved by ACH is statistically significant over the compared baseline methods at the level of $p \leq 0.05$.

C. Experimental Results

Table III shows the mAP results between the baseline methods and their variants based on ACH or its components (DBW and LCR in terms of different metrics with varying hash codes from 8 to 32 bits). The baseline methods enhanced by ACH outperform their original counterparts on two datasets for different code lengths, showcasing a substantial and consistent improvement from ACH. Of particular interest, the COVID-19 Radiograph dataset has a skewed class distribution that may impact the retrieval performance. As shown in Table I, the experiments conducted on two widely used radiological image datasets, demonstrating that ACH achieves significant improvement (0.2%–9%) in mAP over six baseline models, across 8-bit to 32-bit code lengths. Particularly, the methods that address the CB issue (HashNet, ADSH, and TransHash) with ACH achieve average mAP improvements of 5.8%, 1.3%, and 3.1%, respectively. On the other hand, the other methods (DSH, CSQ, and DBDH) with ACH also show reliable average mAP increases of 1.5%, 0.7%, and 5.7%, respectively, highlighting the robustness of ACH to class-imbalanced datasets.

This work presents two key components, namely DBW and LCR, that significantly improve the performance of baseline

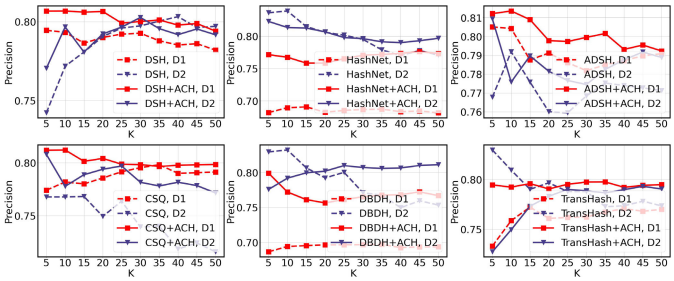

Fig. 7. P@K curves of all the methods on the two datasets at 8-bit (D1: COVID-19 Radiograph; D2: COVID-CT-Dataset).

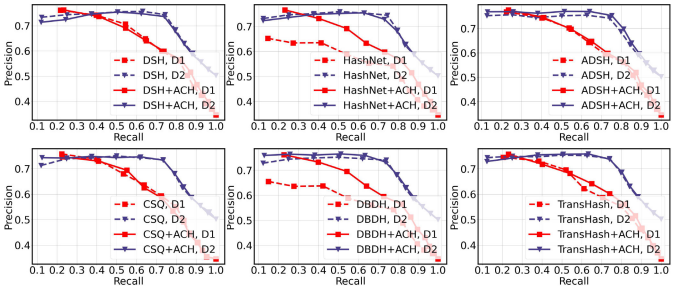

Fig. 8. PR curves of all the methods on the two datasets at 8-bit (D1: COVID-19 Radiograph; D2: COVID-CT-Dataset).

methods as demonstrated in Table III. The results show that both DBW and LCR independently contribute to the improvement of the baseline models. It is also observed that DBW significantly improves the performance of HashNet and TransHash, particularly when LFE is not considered. This highlights the importance of reducing confounder bias between samples in achieving better performance. Additionally, LCR demonstrates its effectiveness in improving the accuracy of lesion positioning, particularly when confounder bias between samples plays a significant role. The combination of DBW and LCR effectively compensates for the limitations of state-of-the-art CBMIR methods and consistently boosts their performance.

The results from the Wilcoxon sign-ranked tests, presented in Table IV, clearly demonstrate that the performance difference between the baseline methods and their ACH variants is statistically significant in terms of mAP, as evidenced by the low p-values. Additionally, Figs. 7 and 8 provide visual representations of the performance improvement achieved by ACH. These figures show the curves of top-K precision (P@K) at 8-bit and PR curves at 8-bit across different datasets. ACH consistently provides stable and comprehensive improvement on the COVID-19 Radiograph dataset and achieves representative results on the COVID-CT dataset for most baseline methods. Particularly, the PR curves demonstrate significant gains for DBDH and HashNet on the COVID-19 Radiograph dataset.

Table V presents the mAP results of ACH and its baselines on the dataset of thermo-graphic IR images. Through analysis, ACH demonstrates superior performance over the compared baseline methods across various hash code lengths. Specifically, ACH outperforms DSH and HashNet in both 8-bit and 16-bit code lengths. Remarkably, it has attained 100% mAP results on the basis of other baseline methods, which may

TABLE III

MAP (%) RESULTS OF THE PROPOSED ACH AND COMPARED BASELINE METHODS ON ALL DATASETS. THE **BEST** AND SECOND BEST RESULTS UNDER DIFFERENT HASH BITS ARE MARKED WITH CORRESPONDING FORMATS. "↑" DENOTES THE IMPROVEMENT TO THE BASELINE RESULTS. "↓" MEANS THE DECREASE TO THE BASELINE RESULTS

Method	COVID-19 Radiograph				COVID-CT-Dataset			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
DSH	74.84	73.85	70.36	72.15	79.48	77.71	77.62	79.20
DSH+DBW(7/8)	74.58	**74.73**	72.85	73.41	**79.88**	79.34	80.38	**80.49**
DSH+LCR(4/8)	74.78	73.97	71.40	71.69	78.96	79.51	79.23	78.52
DSH+ACH(8/8)	**75.08**$_{(\uparrow 0.24)}$	74.64$_{(\uparrow 0.79)}$	**73.57**$_{(\uparrow 3.21)}$	**73.87**$_{(\uparrow 1.72)}$	79.70$_{(\uparrow 0.22)}$	**79.99**$_{(\uparrow 2.28)}$	**80.63**$_{(\uparrow 3.01)}$	79.30$_{(\uparrow 0.10)}$
HashNet	64.89	66.85	73.45	74.85	78.87	79.00	81.33	82.38
HashNet+DBW(7/8)	71.96	**75.80**	76.31	**77.68**	**80.76**	80.91	82.16	82.12
HashNet+LCR(6/8)	67.33	72.27	74.43	75.89	78.71	79.87	81.79	81.80
HashNet+ACH(8/8)	**74.00**$_{(\uparrow 9.11)}$	74.71$_{(\uparrow 7.86)}$	**77.16**$_{(\uparrow 3.71)}$	77.48$_{(\uparrow 2.63)}$	80.68$_{(\uparrow 1.81)}$	**81.58**$_{(\uparrow 2.58)}$	**82.44**$_{(\uparrow 1.11)}$	**82.55**$_{(\uparrow 0.17)}$
ADSH	74.91	76.29	75.92	75.88	79.88	**80.61**	80.66	81.38
ADSH+DBW(7/8)	75.24	76.88	**77.24**	77.64	**81.71**	79.92	81.66	**82.08**
ADSH+LCR(5/8)	75.50	76.28	75.67	75.95	81.54	79.47	81.57	81.73
ADSH+ACH(7/8)	**75.95**$_{(\uparrow 1.04)}$	**76.90**$_{(\uparrow 0.61)}$	77.15$_{(\uparrow 1.23)}$	**77.98**$_{(\uparrow 2.10)}$	81.01$_{(\uparrow 1.13)}$	80.55$_{(\downarrow 0.06)}$	**81.75**$_{(\uparrow 1.09)}$	81.44$_{(\uparrow 0.06)}$
CSQ	73.96	75.53	76.01	76.01	76.06	74.57	76.00	72.75
CSQ+DBW(8/8)	74.15	**76.63**	76.73	76.36	77.29	76.78	78.02	75.60
CSQ+LCR(5/8)	74.04	74.97	75.71	75.98	76.63	75.86	77.38	73.82
CSQ+ACH(8/8)	**74.81**$_{(\uparrow 0.85)}$	75.60$_{(\uparrow 0.07)}$	**76.77**$_{(\uparrow 0.76)}$	**76.92**$_{(\uparrow 0.91)}$	**78.64**$_{(\uparrow 2.58)}$	**78.97**$_{(\uparrow 4.40)}$	**79.97**$_{(\uparrow 3.97)}$	**78.51**$_{(\uparrow 5.76)}$
DBDH	65.10	67.41	74.14	74.65	77.96	79.06	81.22	82.15
DBDH+DBW(7/8)	71.63	**75.59**	76.24	77.41	**81.43**	80.96	**82.22**	82.11
DBDH+LCR(7/8)	70.51	72.22	75.10	76.05	78.92	80.73	81.86	82.08
DBDH+ACH(8/8)	**74.11**$_{(\uparrow 9.01)}$	75.06$_{(\uparrow 7.65)}$	**76.95**$_{(\uparrow 2.81)}$	**77.89**$_{(\uparrow 3.24)}$	81.10$_{(\uparrow 3.14)}$	**81.27**$_{(\uparrow 2.21)}$	81.24$_{(\uparrow 0.02)}$	**82.19**$_{(\uparrow 0.04)}$
TransHash	72.18	74.05	72.88	72.45	79.53	81.07	81.59	81.39
TransHash+DBW(7/8)	**74.43**	**76.03**	75.69	76.09	**81.26**	**81.97**	82.01	81.37
TransHash+LCR(4/8)	72.93	73.07	74.07	73.19	79.50	81.02	81.95	80.93
TransHash+ACH(8/8)	74.27$_{(\uparrow 2.09)}$	74.97$_{(\uparrow 0.92)}$	**77.17**$_{(\uparrow 4.29)}$	**77.45**$_{(\uparrow 5.00)}$	79.76$_{(\uparrow 0.23)}$	81.50$_{(\uparrow 0.43)}$	**82.34**$_{(\uparrow 0.75)}$	**82.19**$_{(\uparrow 0.80)}$

*The brackets represent the proportion that each module wins relative to the baseline method.

TABLE IV

p-VALUES BASED ON MAP OF THE TWO-SIDE WILCOXON SIGN-RANKED TEST FOR COMPARING THE PROPOSED ACH AND BASELINE METHODS ON THE TWO DATASETS. "W," "T," AND "L" ARE THE NUMBER OF WINS, TIES, AND LOSSES, RESPECTIVELY

Method	COVID-19 Radiograph				COVID-CT-Dataset			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
DSH+ACH vs. DSH	4.738e-20	6.904e-4	3.288e-152	2.198e-250	1.633e-06	4.580e-10	1.646e-12	2.632e-07
HashNet+ACH vs. HashNet	6.760e-180	7.720e-78	5.230e-41	6.376e-23	1.358e-12	1.472e-15	1.871e-14	4.126e-2
ADSH+ACH vs. ADSH	5.697e-12	4.787e-28	3.541e-25	1.308e-227	1.207e-13	1.427e-2	9.119e-14	6.164e-08
CSQ+ACH vs. CSQ	1.884e-1	5.441e-16	3.748e-134	5.071e-225	1.278e-13	3.090e-15	3.451e-06	5.974e-15
DBDH+ACH vs. DBDH	3.485e-248	3.976e-61	3.639e-44	6.530e-32	2.730e-14	5.908e-15	1.088e-12	1.047e-2
TransHash+ACH vs. TransHash	4.839e-2	7.319e-1	2.485e-77	1.865e-197	5.860e-13	1.631e-4	9.564e-15	1.860e-08
Summary(W/T/L)	5/0/1	5/0/1	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0

TABLE V

MAP (%) RESULTS OF THE PROPOSED ACH AND COMPARED BASELINE METHODS ON THE DATASET OF THERMO-GRAPHIC IR IMAGES

Method	Dataset of thermo-graphic IR images			
	8 bits	16 bits	24 bits	32 bits
DSH	75.08	86.03	92.47	100.00
DSH+ACH	82.98	92.52	100.00	100.00
HashNet	73.77	86.88	100.00	100.00
HashNet+ACH	73.77	100.00	100.00	100.00
ADSH	100.00	100.00	100.00	100.00
ADSH+ACH	100.00	100.00	100.00	100.00
CSQ	100.00	100.00	100.00	100.00
CSQ+ACH	100.00	100.00	100.00	100.00
DBDH	100.00	100.00	100.00	100.00
DBDH+ACH	100.00	100.00	100.00	100.00
TransHash	100.00	100.00	100.00	100.00
TransHash+ACH	100.00	100.00	100.00	100.00

be attributed to the more pronounced pattern in the thermal imaging dataset. This result demonstrates the adaptability of ACH in domains beyond conventional radiographic contexts.

Based on the comprehensive experimental analysis, we can make the following assessments regarding the capabilities of ACH: 1) ACH can be seamlessly integrated as a plug-and-play module into various baseline hashing methods. Its debiasing properties effectively disentangle confounding factors from their association with labels; 2) ACH exhibits excellent adaptability in addressing challenges such as class imbalance, sensitivity to varying hash code lengths, and variations in optimization strategies; and 3) the DBW and LCR modules, as the two core modules of ACH, play pivotal roles in exploring causal relationships between image representations and labels.

D. Exploration Experiment

The effectiveness of each component in ACH is evaluated through simulation experiments in this section.

1) Extension Application in MRI Images: Given the similarities in appearance and purpose between MRI, X-ray, and CT images, it is of interest to investigate the feasibility of transferring knowledge gained from X-ray or CT images to an MRI-based image retrieval task. To evaluate the effects

TABLE VI

MAP (%) RESULTS OF THE TL FROM THE COVID-19 RADIOGRAPH DATASET AND COVID-CT-DATASET TO THE BRAIN TUMOR DATASET ON HASHNET, CSQ, DBDH, AND THE PROPOSED ACH, RESPECTIVELY

Method	COVID-19 Radiograph → Brain Tumor				COVID-CT-Dataset → Brain Tumor			
	8 bits	16 bits	24 bits	32 bits	8 bits	16 bits	24 bits	32 bits
HashNet w/o TL	35.27	34.74	34.49	33.97	35.27	34.74	34.49	33.97
HashNet w/ TL	43.85	46.11	45.48	41.67	41.50	43.86	47.78	39.57
HashNet+ACH w/ TL(7/8)	45.55	49.19	46.11	44.97	42.31	44.96	47.08	44.34
CSQ w/o TL	36.34	32.58	34.64	33.68	36.34	32.58	34.64	33.68
CSQ w/ TL	56.87	57.71	58.05	59.82	52.28	51.42	54.01	52.71
CSQ+ACH w/ TL(5/8)	57.61	57.95	58.51	59.27	52.41	49.60	54.23	49.25
DBDH w/o TL	34.63	34.69	34.64	34.27	34.63	34.69	34.64	34.27
DBDH w/ TL	42.08	44.44	39.05	39.75	42.42	43.01	40.60	35.12
DBDH+ACH w/ TL(7/8)	46.62	45.69	41.10	38.72	44.71	43.33	42.30	38.47

*The brackets represent the proportion that ACH wins relative to the baseline method.

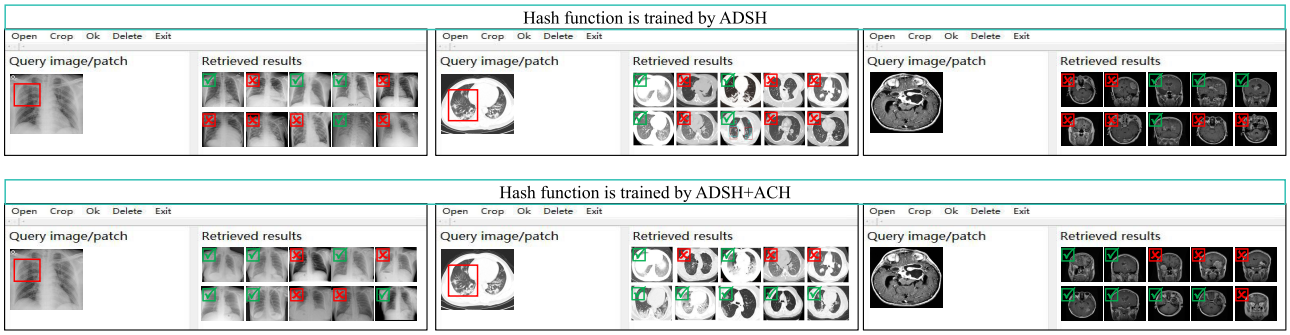

Fig. 9. Examples of the query patches and the top 10 search results are shown. The retrieved results marked with green enclosed ✓are the correct ones, and those marked with red enclosed ✗are the incorrect ones. The upper three rows employ the hash function trained by ADSH, while the bottom three rows utilize the hash function trained by ADSH+ACH. The two applications in the same column use the same query patch.

of LFs and class imbalance, three representative baseline methods are selected for performance comparison, that is, HashNet (with pairwise loss and balanced weight), CSQ (with center loss but without CB), and DBDH (with pairwise loss but without CB). We first pretrain the compared models on either the COVID-19 Radiograph dataset or the COVID-CT dataset and then fine-tune the resultant models on the Brain Tumor dataset for retrieval. The results in Table VI indicate that the models pretrained on the baseline methods or their ACH variants achieve higher retrieval performance compared to results obtained without TL. It can also be seen that the ACH variants achieve higher mAP performance in the TL experiment when compared to their original counterparts, reflecting their good generalization ability. In particular, the pretrained models based on the pairwise loss (HashNet and DBDH) are more suitable for TL from X-ray to MRI image retrieval than the pretrained model based on the center loss method (CSQ).

2) Small-Scale Target Lesion Retrieval: In the clinic, it might be more prudent to identify similar query-relevant candidates with respect to a small-scale image of a target lesion that has been prescribed by doctors for reference [42]. To test this capability, we conduct a simulation experiment on a baseline method and its ACH variant (here ADSH is chosen) on both datasets. First, ten target lesion areas for each class were identified by a professional doctor to construct a toy query set, which consisted of small-scale patches instead of the original images. Then, the toy query set is fed to the model to produce the top 10 results for each query patch. As a result, ADSH can reach the average P@10 of 63.00% and 52.00% on the COVID-19-Radiograph and COVID-CT-Dataset, respectively. When combined with ACH, the average P@10 can be increased by 1.50% and 12.00% on the COVID-19-Radiograph and COVID-CT-Dataset, respectively. Fig. 9 shows two examples of the query patches and the top 10 search results, demonstrating that the ACH variant outperformed ADSH by producing more accurate results on both datasets. In this mode, the search process becomes more efficient, with 7% and 9% decrease in runtime on the COVID-19-Radiograph and COVID-CT-Dataset, respectively. These experimental results show that ACH consistently provides robust improvements for small-scale target lesion retrieval.

3) LCR Versus Lesion Factual Reasoning: To investigate the benefits of counterfactual reasoning in medical image retrieval, factual reasoning [43] as its opponent is also carried out for comparison. As the name implies, lesion factual reasoning (LFR) only concentrates on learning the representation of the lesion area representation while ignoring the impact of nonlesioned areas. In this study, the basic idea of LFR is to generate factual samples by masking the nonlesioned areas in the original images. The main difference between LCR and LFR lies in the part of the original image that is masked during sample generation. For factual sample generation, we just need to rewrite (18) as

$$\mathrm{M}_{1hw} = \begin{cases} 1 & \text{if } w \in \left[\tilde{Y}_P - \frac{s_w}{2}, \tilde{Y}_P + \frac{s_w}{2}\right] \\ & \text{and } h \in \left[\tilde{X}_P - \frac{s_h}{2}, \tilde{X}_P + \frac{s_h}{2}\right] \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

TABLE VII

COMPARISON ON MAP (%) RESULTS FOR LCR AND LESION FACTUAL REASONING. "W," "T," AND "L" ARE THE NUMBER OF WINS, TIES, AND LOSSES, RESPECTIVELY

Method	COVID-19 Radiograph		COVID-CT-Dataset	
	8 bits	32 bits	8 bits	32 bits
DSH+LFR	74.44	71.17	79.23	78.88
DSH+LCR	74.78	71.69	78.96	78.52
DSH+DBW+LFR	74.29	72.96	79.63	79.68
DSH+DBW+LCR	75.08	73.87	79.70	79.30
HashNet+LFR	64.63	74.75	77.84	82.93
HashNet+LCR	67.33	75.89	78.71	81.80
HashNet+DBW+LFR	71.37	77.35	80.01	82.08
HashNet+DBW+LCR	74.00	77.48	80.68	82.55
ADSH+LFR	74.91	75.22	78.93	80.34
ADSH+LCR	75.50	75.95	81.54	81.73
ADSH+DBW+LFR	76.22	77.89	80.59	82.04
ADSH+DBW+LCR	75.95	77.98	81.01	81.44
CSQ+LFR	73.96	75.73	76.31	73.52
CSQ+LCR	74.04	75.98	76.63	73.82
CSQ+DBW+LFR	74.24	76.57	78.02	76.91
CSQ+DBW+LCR	74.81	76.92	78.64	78.51
DBDH+LFR	65.08	74.84	77.52	82.00
DBDH+LCR	70.51	76.05	78.92	82.08
DBDH+DBW+LFR	72.44	77.28	79.73	82.01
DBDH+DBW+LCR	74.11	77.89	81.10	82.19
TransHash+LFR	71.87	71.98	79.50	81.72
TransHash+LCR	72.93	73.19	79.50	80.93
TransHash+DBW+LFR	73.29	75.86	80.73	81.22
TransHash+DBW+LCR	74.27	77.45	79.76	82.19
Summary(W/T/L)	(11/0/1)	(12/0/0)	(9/1/2)	(7/0/5)

The mAP results of both algorithms are shown in Table VII. We can observe that LCR outperforms LFR in most cases for X-ray image retrieval, demonstrating that LCR is more suitable for radiological image retrieval tasks. The inclusion of DBW can improve the performance of both LCR and LFR, indicating that decoupling lesion and confounder information by DBW can enhance lesion detection.

4) Sensitivity to Parameters: There are four key parameters in this work: the width/height ($\tilde{W}/\tilde{H}$) of the sliding window, the width/height ($\tilde{W}^P/\tilde{H}^P$) of the salient area over a key point, the number of key points K in (16), and the scales of masks (s_h, s_w) in (18). To analyze the influences of these parameters, we conduct experiments based on ADSH at the two datasets with a fixed code length of 8 bits. The results of Fig. 10 demonstrate that the performance of ACH is not sensitive to different hyperparameter settings generally. The size of the sliding window and salient area should be modest according to that of the original images. Using more key points cannot necessarily lead to better performance. For counterfactual sample generation, multiscale masking performs better than single-scale masking as expected. Based on the results of Fig. 10, the default hyperparameters of ACH are set to 12, 12, 4, and (2,2), (4,4,), (6,6), (8,8) for $\tilde{W}/\tilde{H}$, $\tilde{W}^P/\tilde{H}^P$, K, and (s_h, s_w), respectively.

5) Adaptability of Real-Time Data: To validate the adaptability of the proposed method for real-time data, we have conducted additional experiments to assess the processing speed of ACH, measured in frames per second (FPS), on three datasets: COVID-19 Radiograph, COVID-CT-Dataset, and Brain Tumor datasets. Our results, presented in Table VIII, demonstrate that ACH consistently achieves FPS exceeding

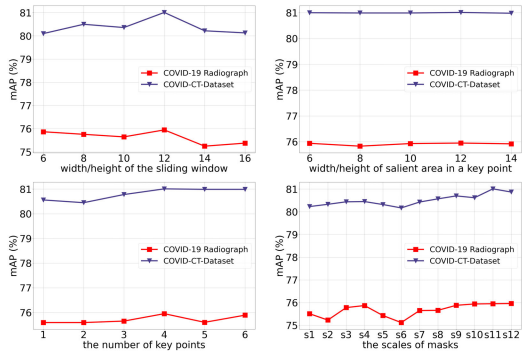

Fig. 10. Analysis of hyperparameters on the two datasets. $s1$–$s12$ refers to {(2,2)}, {(4,4)}, {(6,6)}, {(8,8)}, {(10,10)}, {(12,12)}, {(4,4), (8,8)}, {(6,6), (8,8)}, {(4,4), (6,6), (8,8)}, {(6,6), (8,8), (10,10)}, {(2,2), (4,4), (6,6), (8,8)}, and {(4,4), (6,6), (8,8), (10,10)}.

TABLE VIII

FPS RESULTS FOR THE PROPOSED METHOD ON THREE DATASETS

Data	COVID-19 Radiograph		COVID-CT-Dataset		Brain Tumor	
	8 bits	32 bits	8 bits	32 bits	8 bits	32 bits
FPS	77.52	77.71	77.56	78.38	76.62	76.91

60 FPS across all datasets and hash bit configurations. This performance aligns with the generally accepted threshold for real-time processing [46].

In conclusion, the exploration experiments validate the multifaceted utility and effectiveness of ACH across diverse scenarios, spanning from cross-modal TL to fine-grained lesion retrieval. Additionally, the comparison between lesion factual reasoning and LCR demonstrates the superior benefits of counterfactual reasoning in medical image retrieval. Lastly, the minimal sensitivity of ACH to various hyperparameters enhances its robustness, further solidifying its potential as a reliable solution in medical image retrieval applications.

E. Visualization Analysis

1) Visualization of Salient Maps: To observe how ACH changes the attention of the baseline method, we visualize the patch embedding from the linear projection layer, the class activation map (CAM) [44] derived from the ViT encoder outputs, and gradient-weighted class activation mapping (Grad-CAM) [45] derived from the ViT encoder outputs. The visualization results of patch embedding directly depict the outcomes of intermediate features, while CAM and Grad-CAM offer distinct perspectives for understanding the decision-making processes in the image regions that the model focuses on. As shown in Fig. 11(a), we select ADSH and its ACH variant as a case study. ADSH and its ACH variant can pay attention to the lesion areas more or less for each image. However, we observe that the irrelevant parts, such as skeletons and organs, tend to distract attention from the lesion areas by ADSH. It is demonstrated that such a pseudo-correlation between confounders and lesions could degenerate the performance of existing methods. Differently, ACH tends to avoid this situation and focus more on the pathological regions. Therefore, ACH can effectively alleviate the issue of confounder bias to achieve better performance.

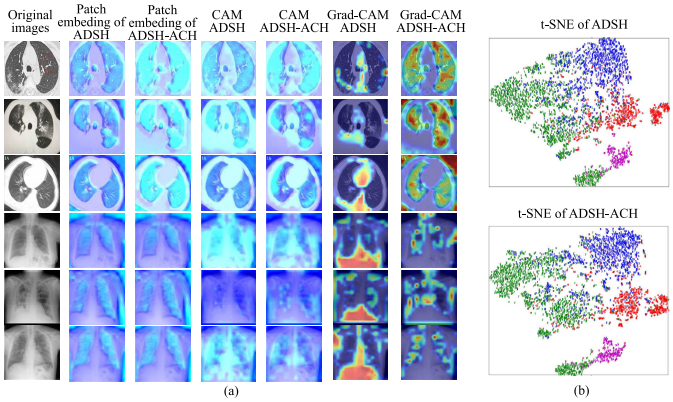

Fig. 11. (a) Visualization of salient maps for the patch embedding, CAM, and Grad-CAM. The bright blue/red/yellow parts are the areas that the methods focus on. (b) Visualization of t-SNE for ADSH and its variant ACH on the COVID-19 Radiograph test dataset.

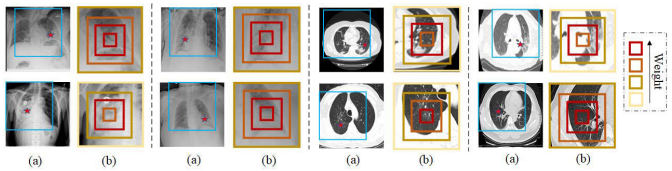

Fig. 12. Lesion area detection via multiscale masking in two stages. Images of lines (a) are original images and images of lines (b) are lesion areas cropped from original images based on the lesion points. The areas with blue boxes in lines (a) are locations of sliding windows in the first searching stage. Red stars are lesion points detected at the second searching stage. Boxes in line (b) are counterfactual masks with different scales, and the importance of each scale is presented with different colors.

2) Visualization of t-SNE: We employ t-SNE testing to gain insights into the reliability of features extracted by ACH. This experiment is conducted on the test set of COVID-19 Radiograph. The visualization of t-SNE is illustrated in Fig. 11(b), ADSH exhibits unclear classification boundaries across the three categories, while the data structure obtained by ACH successfully separates different categories. This demonstrates that ACH excels in discriminating between challenging categories.

3) Lesion Area Detection via Multiscale Masking: Fig. 12 showcases how to detect lesion area by our proposed method in two stages. In the first stage, the C2F-LP module allows to locate the lesion area by a sliding window (colored in blue). In the second stage, the lesion key points are accurately identified by counterfactual reasoning. The intensity of the color represents different weights of multiscale masks over the key point. It is shown that such a coarse-to-fine searching strategy makes it more effective in detecting possible lesion areas with flexible multiscale masking. The lesion areas masked with high weights are also helpful to assist doctors in rapidly locating the target lesion.

Through the visualization experiments, we draw the following conclusions regarding the interpretability of ACH: 1) ACH prioritizes regions of interest based on pathology, unaffected by confounding factors, providing physicians with reliable visualizations and 2) the incorporation of multiscale masking demonstrates the precision of the two-stage lesion detection module C2F-LP.

VI. Conclusion

This article proposes a novel ACH method for radiological image retrieval by alleviating confounder bias. Specifically, ACH captures the NDE of lesions on the generated hash code, which is accomplished by addressing both the original similarity measurement problem and the counterfactual similarity measurement problem simultaneously. A sample DBW method based on backdoor adjustment is proposed to disentangle the relationship between lesions and confounding factors. Furthermore, the proposed LCR module accurately detects lesions, which allows for generating counterfactual samples with masked lesions. The exhaustive experimentation across four real-world datasets has substantiated the efficacy and dependability of ACH. The experiment outcomes have brought to light specific advantages that deserve particular attention.

1) ACH can be seamlessly integrated as a plug-and-play modular enhancement to baseline hashing methods. This integration is designed to enhance retrieval accuracy through causal inference without adding to the time or space complexity during online inference.
2) ACH is powered by the DBW and LCR modules with the ability to separate lesion features from confounding factors using backdoor adjustment and reinforcing its robustness via counterfactual reasoning.
3) ACH allows physicians to perform more fine-grained retrieval on regions of interest, thus aligning retrieval processes more closely with clinical intent.
4) The accurate visualizations of ACH offer explanatory insights to facilitate a better understanding and interpretation of the diagnostic decision-making process.

While ACH has achieved significant advancements, we must carefully reflect on its limitations. The counterfactual reasoning process in ACH necessitates dual access to the backbone structure during the training phase, which leads to longer training durations. Additionally, employing a mask-based approach for generating counterfactual samples may compromise the integrity of organ structures in medical images.

Based on the aforementioned limitations of ACH, our future endeavors aim to develop more sophisticated techniques for precise counterfactual sample generation, which help clinicians predict how changes in treatment strategies might affect patient outcomes. Additionally, with the rapid advancement of large-scale medical models, we plan to investigate how these extensive knowledge bases can be integrated into our framework to enhance causal inference capabilities.

References

[1] J. Wu et al., "Radiological tumour classification across imaging modality and histology," *Nature Mach. Intell.*, vol. 3, no. 9, pp. 787–798, Aug. 2021.
[2] E. Yang et al., "Deep Bayesian hashing with center prior for multimodal neuroimage retrieval," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 503–513, Feb. 2021.
[3] Y. Hu et al., "Source free semi-supervised transfer learning for diagnosis of mental disorders on fMRI scans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13778–13795, Nov. 2023.
[4] Z.-A. Huang et al., "Federated multi-task learning for joint diagnosis of multiple mental disorders on MRI scans," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 4, pp. 1137–1149, Apr. 2023.

[5] F. Z. Benyelles, A. Sekkal, and N. Settouti, "Content based COVID-19 chest X-ray and CT images retrieval framework using stacked auto-encoders," in *Proc. 2nd Int. Workshop Hum.-Centric Smart Environ. Health Well-Being (IHSH)*, Feb. 2021, pp. 119–124.

[6] Y. Zheng et al., "Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102308.

[7] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fMRI data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 10591–10605, Aug. 2024.

[8] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7627–7641, Jun. 2024.

[9] V. Kumar, V. Tripathi, and B. Pant, "Content based surgical video retrieval via multi-deep features fusion," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2021, pp. 1–5.

[10] J. Pearl et al., *Causality: Models, Reasoning, and Inference*, vol. 19. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[11] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. VIDB Conf.*, 1999, vol. 99, no. 6, pp. 518–529.

[12] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[13] T. Hoang, T.-T. Do, T. V. Nguyen, and N.-M. Cheung, "Multimodal mutual information maximization: A novel approach for unsupervised deep cross-modal hashing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6289–6302, Sep. 2023.

[14] M. Yu, L. Liu, and L. Shao, "Binary set embedding for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2899–2910, Dec. 2017.

[15] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2064–2072.

[16] Q.-Y. Jiang and W.-J. Li, "Asymmetric deep supervised hashing," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Apr. 2018, vol. 32, no. 1, pp. 1–5.

[17] L. Jin, K. Li, Z. Li, F. Xiao, G.-J. Qi, and J. Tang, "Deep semantic-preserving ordinal hashing for cross-modal similarity search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1429–1440, May 2018.

[18] Q. Qin, K. Xie, W. Zhang, C. Wang, and L. Huang, "Deep neighborhood structure-preserving hashing for large-scale image retrieval," *IEEE Trans. Multimedia*, vol. 26, pp. 1881–1893, 2023.

[19] L. Yuan et al., "Central similarity quantization for efficient image and video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3083–3092.

[20] L. Wang, Y. Pan, C. Liu, H. Lai, J. Yin, and Y. Liu, "Deep hashing with minimal-distance-separated hash centers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23455–23464.

[21] Z.-A. Huang, R. Liu, Z. Zhu, and K. C. Tan, "Multitask learning for joint diagnosis of multiple mental disorders in resting-state fMRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8161–8175, Jun. 2024.

[22] M. Yang, J. Xu, W. Ding, and Y. Liu, "FedHAP: Federated hashing with global prototypes for cross-silo retrieval," *IEEE Trans. Parallel Distrib. Syst.*, vol. 35, no. 4, pp. 592–603, Apr. 2024.

[23] W. Tan, L. Zhu, J. Li, H. Zhang, and J. Han, "Teacher-student learning: Efficient hierarchical message aggregation hashing for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 25, pp. 4520–4532, 2023.

[24] J. Fang, H. Fu, D. Zeng, X. Yan, Y. Yan, and J. Liu, "Combating ambiguity for hash-code learning in medical instance retrieval," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3943–3954, Oct. 2021.

[25] Y. Gu, K. Vyas, M. Shen, J. Yang, and G. Yang, "Deep graph-based multimodal feature embedding for endomicroscopy image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 481–492, Feb. 2021.

[26] Y. Zheng et al., "Diagnostic regions attention network (DRA-Net) for histopathology WSI recommendation and retrieval," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 1090–1103, Mar. 2021.

[27] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Mach. Intell.*, vol. 3, no. 7, pp. 610–619, May 2021.

[28] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, "C-CAM: Causal CAM for weakly supervised semantic segmentation on medical image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11676–11685.

[29] J. J. Thiagarajan, K. Thopalli, D. Rajan, and P. Turaga, "Training calibration-based counterfactual explainers for deep learning models in medical image analysis," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, Jan. 2022.

[30] D. Lenis, D. Major, M. Wimmer, A. Berg, G. Sluiter, and K. Bühler, "Domain aware medical image classifier interpretation by counterfactual impact analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*, Jul. 2020, pp. 315–325.

[31] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 61–71.

[32] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–10.

[33] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep learning to hash by continuation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5608–5617.

[34] X. Zheng, Y. Zhang, and X. Lu, "Deep balanced discrete hashing for image retrieval," *Neurocomputing*, vol. 403, pp. 224–236, Aug. 2020.

[35] Y. Chen, S. Zhang, F. Liu, Z. Chang, M. Ye, and Z. Qi, "TransHash: Transformer-based Hamming hashing for efficient image retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 127–136.

[36] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[37] M. E. H. Chowdhury et al., "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[38] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: A CT scan dataset about COVID-19," 2020, *arXiv:2003.13865*.

[39] S. Bhuvaji, A. Kadam, P. Bhumkar, S. Dedge, and S. Kanchan. (2020). *Brain Tumor Classification (MRI)*. Kaggle, doi: 10.34740/KAGGLE/DSV/1183165. [Online]. Available: https://www.kaggle.com/dsv/1183165

[40] M. Najafi, Y. Baleghi, S. A. Gholamian, and S. M. Mirimani, "Fault diagnosis of electrical equipment through thermal imaging and interpretable machine learning applied on a newly-introduced dataset," in *Proc. 6th Iranian Conf. Signal Process. Intell. Syst. (ICSPIS)*, Dec. 2020, pp. 1–7.

[41] A. Glowacz, "Thermographic fault diagnosis of electrical faults of commutator and induction motors," *Eng. Appl. Artif. Intell.*, vol. 121, May 2023, Art. no. 105962.

[42] Z. Ye, Y. Zhang, Y. Wang, Z. Huang, and B. Song, "Chest CT manifestations of new coronavirus disease 2019 (COVID-19): A pictorial review," *Eur. Radiol.*, vol. 30, no. 8, pp. 4381–4389, Aug. 2020.

[43] C.-H. Chang, G. A. Adam, and A. Goldenberg, "Towards robust classification model by counterfactual and invariant data generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15212–15221.

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[46] P. Cížek and J. Faigl, "Real-time FPGA-based detection of speeded-up robust features using separable convolution," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1155–1163, Mar. 2018.

Yajie Zhang received the M.Sc. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2022. She is currently pursuing the Ph.D. degree with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, SAR.

She was a Research Assistant with the City University of Hong Kong (Dongguan), Dongguan, China. Her research interests include artificial intelligence, medical image analysis, and causal inference.

Yao Hu (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2024.

He is currently a Post-Doctoral Fellow with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, SAR. His research interests include domain adaptation, federated learning, evolutionary transfer optimization, and applied deep learning for medical image analysis.

Chengjun Cai (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2021.

He is currently an Associate Professor at the City University of Hong Kong (Dongguan), Dongguan, China. His research interests include applied cryptography, data security and privacy, and blockchain.

Dr. Cai is a member of ACM.

Yu-An Huang (Member, IEEE) received the Ph.D. degree in computer science from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR, in 2020.

He is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China. His research interests include computational biology, deep learning, and data mining.

Zhi-An Huang (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, SAR, in 2021.

He is currently an Assistant Professor with the City University of Hong Kong (Dongguan), Dongguan, China. He has authored or co-authored more than 30 papers in esteemed journals and conference proceedings. His research interests include AI in healthcare, machine learning, bioinformatics, and medical imaging analysis.

Dr. Huang is currently an Associate Editor of IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS.

Kay Chen Tan (Fellow, IEEE) received the B.Eng. (Hons.) and Ph.D. degrees from the University of Glasgow, Glasgow, U.K., in 1994 and 1997, respectively.

He is currently a Chair Professor of computational intelligence with the Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, SAR. He is also an Honorary Professor with the University of Nottingham, Nottingham, U.K., and the Chief Coeditor of the Springer Book Series on *Machine Learning: Foundations, Methodologies, and Applications.*

Dr. Tan was the Editor-in-Chief of IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and currently serves as an editorial board member for more than ten journals.