Information Sciences 637 (2023) 118938

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Content-based medical image retrieval with opponent class adaptive margin loss

Şaban Öztürk [a,b,c,*], Emin Çelik [a,b], Tolga Çukur [a,b,d]

[a] Department of Electrical and Electronics Engineering, Bilkent University, TR-06800 Ankara, Turkey
[b] National Magnetic Resonance Research Center, Bilkent University, TR-06800 Ankara, Turkey
[c] Department of Electrical and Electronics Engineering, Amasya University, TR-05001 Amasya, Turkey
[d] Neuroscience Program, Sabuncu Brain Research Center, Bilkent University, TR-06800 Ankara, Turkey

ARTICLE INFO

Keywords:
CBIR
Medical image retrieval
Triplet
Representational learning
Hashing

ABSTRACT

The increasing utilization of medical imaging technology with digital storage capabilities has facilitated the compilation of large-scale data repositories. Fast access to image samples with similar appearance to suspected cases in these repositories can help establish a consulting system for healthcare professionals, and improve diagnostic procedures while minimizing processing delays. However, manual querying of large repositories is labor intensive. Content-based image retrieval (CBIR) offers an automated solution based on quantitative assessment of image similarity based on image features in a latent space. Since conventional methods based on handcrafted features typically show poor generalization performance, learning-based CBIR methods have received attention recently. A common framework in this domain involves classifier-guided models that are trained to detect different image classes. Similarity assessments are then performed on the features captured by the intermediate stages of the trained models. While classifier-guided methods are powerful in inter-class discrimination, they are suboptimally sensitive to within-class differences in image features. An alternative framework instead performs task-agnostic training to learn an embedding space that enforces the representational discriminability of images. Within this representational-learning framework, a powerful method is triplet-wise learning that addresses the deficiencies of point-wise and pair-wise learning in characterizing the similarity relationships between image classes. However, the traditional triplet loss enforces separation between only a subset of image samples within the triplet via a manually-set constant margin value, so it can lead to suboptimal segregation of opponent classes and limited generalization performance. To address these limitations, we introduce a triplet-learning method for automated querying of medical image repositories based on a novel Opponent Class Adaptive Margin (*OCAM*) loss. To maintain optimally discriminative representations, *OCAM* considers relationships among all image pairs within the triplet and utilizes an adaptive margin value that is automatically selected per dataset and during the course of training iterations. CBIR performance of *OCAM* is compared against state-of-the-art loss functions for representational learning on three public databases (gastrointestinal disease, skin lesion, lung disease). On average, *OCAM* shows an mAP performance of 86.30% in the KVASIR dataset, 70.30% in the ISIC 2019 dataset, and 85.57% in the X-RAY dataset. Comprehensive experiments in each application domain demonstrate the superior performance of *OCAM* against competing triplet-wise methods at 1.52%, classifier-guided methods at 2.29%, and non-triplet representational-learning methods at 4.56%.

* Corresponding author at: Department of Electrical and Electronics Engineering, Bilkent University, TR-06800 Ankara, Turkey.
E-mail addresses: saban.ozturk@amasya.edu.tr (Ş. Öztürk), cukur@ee.bilkent.edu.tr (T. Çukur).

https://doi.org/10.1016/j.ins.2023.118938
Received 27 November 2022; Received in revised form 5 April 2023; Accepted 11 April 2023
Available online 13 April 2023

1. Introduction

With the tremendous increase in the volume of medical imaging data accumulated in healthcare institutions due to the increasing availability of imaging devices, promising opportunities have arisen in establishing image-based consulting systems for healthcare professionals. Yet, a key component of such systems is the ability to efficiently search representative images of suspected diseases within large image repositories. A powerful solution to address this challenge is content-based image retrieval (CBIR), which can automate the search for images similar to a query by assessing the similarity of their image features. By providing automated access to medical images that are visually similar to a query image, CBIR systems can aid in the solution of clinical problems such as grading of disease progression, diagnosing multiple concurrent diseases, cross-organ disease assessment, and counseling of medical trainees. Automated retrieval is characteristically performed by comparing the visual contents of the query image against candidate images in a large image repository. Unfortunately, conventional CBIR methods assess the similarity of visual contents in terms of hand-crafted local and global image features [1]. Since these approaches often rely on a compact set of color and texture features, they provide an incomplete characterization of the image distribution and show suboptimal generalization to new image samples [2].

The limitations of conventional methods have recently motivated the adoption of learning-based CBIR methods with the advent of deep neural network (DNN) architectures for data-driven capture of comprehensive image features. A prevalent framework in this domain employs classifier-guided DNNs trained to detect the image classes via cross-entropy losses [3]. The feature maps extracted in intermediate stages of the trained DNNs are then used to conduct similarity assessments between images. While this framework is powerful in inter-class discriminations, classifier-guided DNNs lead to suboptimal intra-class discriminability as that information is not relevant to the classification task [4]. An alternative framework involves task-agnostic training of DNNs to capture a latent embedding space where images are highly discriminable [5]. In this representational-learning framework, each image is represented as a dense embedding vector, and similarity between two images is computed based on their respective embedding vectors. To do this, an embedding objective is used that enforces the assignment of comparable vectors to visually similar images, and disparate vectors to visually dissimilar images [6]. Consequently, retrieval is performed by recollecting the set of images that are closest to the query image in the latent space. Since representational learning does not exclusively focus on class discrimination, it can capture discernible representations for inter-class and intra-class images.

Prominent embedding objectives to learn representations that become more discernible with growing dissimilarity between images include pair-wise, and triplet-wise methods. Pair-wise methods process images in pairs to enable similarity assessments based on contrastive loss [7]. However, these methods have relatively limited training efficiency as they only consider pairs of either similar or dissimilar images, and they can produce weakly discernible representations for similar images. To address these limitations, triplet-wise methods process images in sets of three with anchor (*A*), positive (*P*), and negative (*N*) samples. A triplet loss is commonly used to compare the *A-P* distance between images of the same class (i.e., similar images) against the *A-N* distance between images of different classes (i.e., dissimilar images) [8]. Traditional triplet-wise methods enforce the difference between *A-P* and *A-N* distances to remain above a constant margin value. Several modifications have also been introduced to the loss function including the linear or non-linear weighting of *A-P* and *A-N* distances in the triplet loss [9,10], and the addition of regularization terms based on the *A-P* distance itself to emphasize its contribution [11,12]. Yet, similar to the traditional formulation, these methods do not place any explicit constraints on the *P-N* distance [8], which can lower the segregation between *P* and *N* samples from opponent classes [13]. Furthermore, previous methods pervasively use constant margin values that can elicit suboptimal performance since the ideal margin varies across datasets and across the training iterations for a given dataset [14].

This work introduces an Opponent Class Adaptive Margin (*OCAM*) loss to improve triplet-wise representational learning for CBIR tasks. Addressing two main limitations of the traditional triplet formulation, *OCAM* incorporates the *P-N* distance to enforce better segregation between opponent classes, and it leverages an adaptive margin determined based on the current segregation between the opponent classes. This allows *OCAM* to learn more discriminative embedding vectors for medical imaging that can facilitate subsequent image retrieval. Demonstrations are provided on gastrointestinal, skin, and lung images based on Euclidean and Hamming retrieval codes. Our experiments indicate that *OCAM* exhibits superior performance against competing baselines. Our main contributions are summarized below:

- *OCAM* leverages an adaptive margin between *A-P* and *A-N* distances to improve conformity to the image distribution per dataset, without necessitating manual intervention.
- *OCAM* incorporates the *P-N* distance in the embedding objective to enhance the discernibility of opponent image classes in the latent space.
- Superior retrieval performance is obtained in various anatomies with the *OCAM*-based CBIR method.

2. Related work

An early approach for querying medical image repositories is text-based image retrieval (TBIR) where image contents are annotated manually by human observers, and a query search is conducted based on the text annotations. Due to the involvement of the human factor, TBIR systems are labor-intensive and susceptible to subjective annotations that vary across observers, rendering them impractical for large repositories [15]. CBIR systems offer an efficient alternative where retrieval is conducted based on the similarity of visual features in medical images instead of text annotations [16]. In CBIR, image contents are represented via features in a latent space, similarity scores between a query image and candidate images from the repository are computed based on these features, and

retrieval is performed by recollecting the candidates with the highest scores [3]. CBIR systems' performance inevitably depends on the representational power of the features used for measuring image similarity [17]. To this end, feature extraction for image retrieval tasks has been an active research area in recent years. Conventional methods in this domain rely on expert knowledge to define hand-crafted features. These include both global features related to color, shape and texture [18] and local features as extracted by various transforms [19]. While these hand-crafted features are relatively easy to derive, conventional methods commonly leverage a compact set of features that are insufficient in providing a comprehensive characterization of the image distribution [16].

Learning-based methods based on DNNs have recently gained traction in CBIR due to their remarkable capabilities in capturing descriptive image features. The first group of methods in this domain follows a classifier-guided framework where a DNN is trained to perform image classification [20]. Training is commonly conducted via a cross-entropy loss for these methods, although tree-based, logistic and graph-based losses have also been proposed [3,21]. Afterward, the feature maps in later layers of the trained DNN are used to assess image similarity. Pooling of features across both intermediate and later DNN layers [22], across multi-scale DNNs [23], and across randomly sampled image patches [24] have been proposed to improve feature diversity at the expense of computational burden. Feature selection methods to lower the dimensionality of the latent representations have later been introduced to alleviate this burden [25]. Due to their data-driven capture of image features, classifier-guided methods have enabled leaps in CBIR performance over conventional methods. However, these methods are primarily trained for a classification task that prioritizes inter-class discrimination, so representations of images within a given class may be less discernible. In turn, classifier-guided methods are amenable to selecting candidate images that are not ideally similar to the query image.

An alternative group of methods instead follow a representational-learning framework where a DNN is trained to identify image embeddings [8]. The embedding objective serves to promote visually similar images to have proximate embedding vectors and visually dissimilar images to have distant embedding vectors. To do this, pair-wise and triplet-wise learning are two common approaches for constructing the embedding objective. In pair-wise learning, a pair of images are processed with a pair of DNNs with matching model weights. The resultant feature maps from the DNNs are then compared to compute a contrastive loss [26]. Pair-wise methods are reported to elevate retrieval performance over classifier-guided methods [7]. However, the contrastive loss is computed based on either a pair of similar or dissimilar images per iteration, which can limit learning efficiency. In addition, the traditional contrastive loss enforces the distance between dissimilar images to exceed a margin value, albeit it does not consider the distance between similar images. For improved learning, triplet-wise methods process three images (*A*, *P*, and *N*) per iteration with *A* and *P* sampled from an opponent class to *N* [8]. As such, the triplet loss enforces the inter-class distance to be greater than the intra-class distance by a constant margin value. While demonstrated to be powerful, conventional triplet-wise methods still face some challenges in representational learning. First, the traditional triplet loss ignores the *P-N* distance in its formulation, resulting in suboptimal segregation between opponent image classes. Second, it uses a user-determined scalar as the margin value that does not attune itself to specific datasets, and that remains constant during the training, resulting in poor generalization. To address these challenges, *OCAM* explicitly incorporates the *P-N* distance and utilizes an adaptive margin value.

Several recent studies have attempted to improve triplet learning by adapting margin values. In [10], a cross-validated search is proposed for tuning the relative weighting of a constant margin per dataset. However, this approach still retains a fixed margin across the entire training process. In [11,12,14], a global constant margin is prescribed by the user, and its relative weighting is adjusted via penalty terms included in the triplet loss. While this approach introduces a degree of flexibility in changing the effective margin across training iterations, it still requires manual selection of the global margin. In [27], an adaptive margin is proposed that is determined by ground-truth *A-P* and *A-N* distances between triplets of image samples. To obtain the ground-truth distances, human observers rate images from distinct classes of visual objects. This labor-intensive rating is difficult to adopt for large training sets, and visual ratings might not be sufficiently distinctive for medical images where differences among separate disease types are relatively modest. In [28], an adaptive margin approach is introduced that is determined per mini-batch of triplet samples. The margin value is defined as an affine function of a semantic similarity term between *A* and *N* samples and a constant base margin term. This adaptive approach ignores the *P* samples while computing the margin, and it still requires manual selection of the base margin term so it is not fully automated. For cases when explicit class information is available on each image, [29] proposes a method to select a class-specific adaptive margin per mini-batch. This method is not applicable when only binary similarity-dissimilarity information is available, and it requires the manual selection of hyperparameters to update the margin across training iterations. [30] proposes to start the training with a small initial margin and to gradually increase it across training iterations by a constant update multiplier as the proportion of *N* samples that can be successfully separated exceeds a threshold ratio. While promising results have been reported, this approach still requires manual selection of the initial margin, the threshold ratio, and the constant update multiplier. In contrast to previous methods, the proposed adaptive margin in *OCAM* is automatically attuned for each dataset and triplet sample across the training iterations. Unlike methods that involve the manual selection of hyperparameters to set the weighting or range of margin values, *OCAM* does not require any manual parameter selection. Unlike methods that rely on external image ratings, *OCAM* computes the margin value based on the *P-N* distance of triplet samples in the learned embedding space. Unlike methods that leverage explicit class information, *OCAM* only relies on binary similarity-dissimilarity labels. Finally, unlike methods that require user-defined parameters, *OCAM* performs fully-automated selection of the adaptive margin.

A recent study on person re-identification has proposed to explicitly integrate the *P-N* distance into the triplet loss to improve performance [13]. Considering not only the *A-P* and *A-N* distances but also the *P-N* distance can help lower the segregation between image samples from opponent classes. Our proposed method differs from [13] in the following aspects. On the one hand, [13] uses a constant margin value as in the traditional triplet loss. This can result in performance degradation as the ideal margin varies across datasets and training iterations due to changing inter-class separation. While manual tuning might partly alleviate the degradations across datasets, adapting the margin value across training iterations is challenging without an automated approach. On the other hand,

unlike [13], *OCAM* also embeds the added *P-N* distance term into its adaptive margin value that is designed to vary inversely with inter-class separation. To avoid the need for manual tuning, the margin is defined using *P-N* distance as a proxy for inter-class separation, so it is automatically adjusted during the course of iterations and per dataset.

3. Theory

3.1. Mathematical preliminaries

Given a single query image, CBIR methods aim to retrieve a finite subset of Z images from a repository with high similarity in visual content to the query. Let's consider a repository $D=\{X,Y\}^K$ consisting of K medical images $X=\{x_1,x_2,\ldots,x_K\} \in R^{dxK}$, where $x_k \in R^{dx}{}^{(1\leq k\leq K)}$ is the kth sample of X,$Y=\{y_1,y_2,\ldots,y_K\}$ represents image labels, and d denotes the image dimensionality. For assessing the similarity of visual content, the images are typically mapped onto a latent space via a projection function $\Omega{:}x_k \rightarrow E_k$, where Ω transforms the original image space to the latent space, and the set of embedding vectors for the images are given as $E=\{E_1,E_2,\ldots,E_K\} \in R^S$. To retrieve the most similar images from the repository, a search for the Z nearest neighbors (NN) to the query is required [30]:

$$\mathrm{NN}(x_q,Z) = \{f(E_q,E_k) \in F : |F \cap (-\infty, f(E_q,E_k))| < Z\} \tag{1}$$

where E_q denotes the embedding of the query image, f is a distance metric, and F denotes the set of distances between the embeddings of repository images and the query image. It is possible to conduct the search by ranking images according to the Euclidean distance of their continuous embedding codes. While search based on continuous codes can be more sensitive, it also introduces a computational burden for large image repositories. For improved search efficiency, a binary hash code can be generated for each image based on its embedding, $B=\{b_1,b_2,\ldots,b_K\}\in\{-1,+1\}^{SxK}$ [31] via a binarization operation:

$$b_k = sgn(E_k) = \begin{cases} +1, & E_k^i \geqslant 0 \\ -1, & E_k^i < 0 \end{cases}, \; i = 1,2,\ldots,S \tag{2}$$

Images can be ranked based on the similarity of their hash codes using the Hamming distance. While hash codes improve time efficiency, their retrieval performance is inevitably dependent on the representation power of the codes. Therefore, DNN methods are commonly adopted to obtain hash codes with high representation power to capture the underlying dense embedding vectors for medical images.

Triplet-wise learning is considered one of the most effective approaches for capturing latent representations of images. The traditional triplet loss (*Triplet*) for representational learning samples a set of three images (x_A, x_P, and x_N) from the repository, as illustrated in Fig. 1. Assuming access to information regarding whether any pair of images belong to the same image class, once x_A is initially selected, x_P is drawn from the same class, whereas x_N is drawn from the opponent class. Embeddings of the images in the triplet are computed via Ω, $\Omega{:}x_A,x_P,x_N \rightarrow E_A,E_P,E_N$, $E \in R^S$. The loss is then expressed as:

$$Triplet(E_A,E_P,E_N) = \max(0, f(E_A,E_P) - f(E_A,E_N) + \alpha) \tag{3}$$

where α represents the margin parameter. A first limitation of the traditional formulation is that, for a random selection of the image triplet, it is possible that $f(E_A,E_P) \geq f(E_P,E_N)$ even if the condition in Eq. (3) is satisfied as $f(E_A,E_P) + \alpha \leq f(E_A,E_N)$. This lack of explicit control over $f(E_P,E_N)$ can lower the discernability of learned embeddings E_P, E_N. A second limitation is that the margin value α requires manual tuning for each dataset, which can be labor-intensive. Moreover, the margin value remains constant throughout the entire training process, which may result in suboptimal performance for specific stages of training.

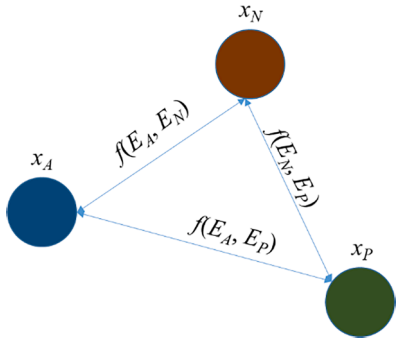

Fig. 1. To define a triplet loss, three image samples are selected. These include an anchor sample (x_A, *in blue color*), a positive sample similar to the anchor (x_P, *in green color*), and a negative sample dissimilar to the anchor (x_N, *in red color*). Links between image samples illustrate their relative distances. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

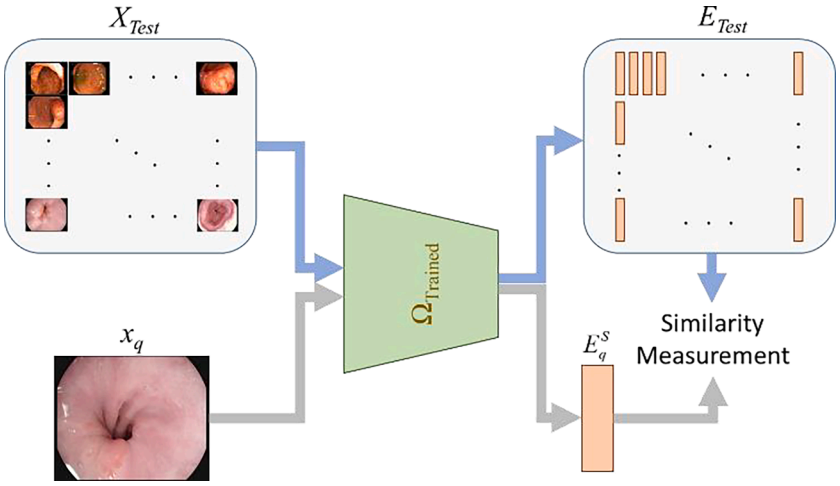

Fig. 2. For CBIR, given a query image x_q, E_q is computed based on a trained network $\Omega_{Trained}$ and compared against the embedding vectors in the repository E_{Test} to retrieve similar images. Candidate images from the repository can be ranked based on the Euclidean distance of continuous embedding codes for the accuracy or based on Hamming distance of binarized hash codes for efficiency.

3.2. Ocam

We introduce *OCAM,* a novel triplet-wise method for medical image retrieval that improves image representation discernibility and generalization performance by incorporating two technical improvements over traditional triplet learning. Traditional triplet learning ignores the *P-N* distance, $f(E_P,E_N)$, so it can yield insufficient segregation between opponent image classes. To address this issue, *OCAM* explicitly incorporates $f(E_P,E_N)$ in the loss function as inspired by a recent computer vision study on person re-identification [13]. Note that there are two samples in each image triplet belonging to the positive class, so weighting the distance terms equally can introduce unwanted biases towards the positive class. To avoid potential biases, a balanced weighting is instead adopted here: ($f(E_A,E_P)$-($f(E_A,E_N)$ + $f(E_P,E_N)$)/2). As such, *OCAM* improves inter-class segregation by extending the *A-P* distance over the *A-N* distance and maintaining a relatively large *P-N* distance. Furthermore, utilizing a static margin can result in suboptimal performance since $f(E_A,E_P)$-$f(E_A,E_N)$ inherently changes across the training process. In particular, an α value suited for early iterations where ($f(E_A,E_P)$-($f(E_A,E_N)$ + $f(E_P,E_N)$)/2) is relatively small will be rendered suboptimal towards later iterations as inter-class segregation increases, slowing down the learning process. To address this limitation, *OCAM* introduces an adaptive margin value inspired by the recent success of adaptive methods in classification tasks [14,27]. Here we propose to leverage a margin value $\alpha_{adaptive}$=(1-$f(E_P,E_N)$)/2 that is a function of $f(E_P, E_N)$ so as to improve the *P-N* separation, avoiding the need for introducing a user-controlled parameter.

Taken together, these design elements lead to the following formulation in *OCAM*:

$$OCAM(E_A,E_P,E_N) = \max(0, f(E_A,E_P) - \frac{(f(E_A,E_N)+f(E_P,E_N))}{2} + \frac{(1-f(E_P,E_N))}{2}) \tag{4}$$

which can be further simplified as:

$$OCAM(E_A,E_P,E_N) = \max(0, f(E_A,E_P) - \left(\frac{f(E_A,E_N)+2f(E_P,E_N)-1}{2}\right)) \tag{5}$$

To learn embeddings with strong representation capability, a cosine distance measure that ranges in [0,1] is used [26,31]:

$$f(E_i,E_j) = \frac{(1-\cos(E_i,E_j))}{2} \tag{6}$$

Then Ω parameters are updated using Adam optimizer according to *OCAM* value. Alg. 1 outlines the training procedures for *OCAM*:

Algorithm 1 *OCAM* training

Input:
$D = \{X,Y\}^K$: training repository consisting of K medical images
$X = \{x_1,x_2,\ldots,x_K\}$: images in the repository
$Y = \{y_1,y_2,\ldots,y_K\}$: labels in the repository
t: max iteration number

(*continued on next page*)

(*continued*)

Algorithm 1 *OCAM* training

n: number of local iteration
Ω_θ: DNN
θ: parameters of DNN
Opt(): Adam optimizer that computes parameter updates
Output:
Ω_{θ^*}: trained DNN
Initialize parameters:
Random initialization of θ
for $n = 1$: t **do**
  Randomly select triplets for iteration n, $\{x_A^k, x_P^k, x_N^k\}_{k=1}^K$
  Create embeddings, $\{E_A^k, E_P^k, E_N^k\}_{k=1}^K = \Omega_\theta\{x_A^k, x_P^k, x_N^k\}_{k=1}^K$
  Calculate $OCAM(E_A^k, E_P^k, E_N^k)$
  Update $\theta, \theta \leftarrow \theta - Opt(\nabla_\theta, OCAM(E_A^k, E_P^k, E_N^k))$
return: $\theta^* = \theta$

Following the training of a neural network Ω_{Trained} according to the loss in Eq. (5), inference can be performed for a query image x_q and a test repository $D_{Test}=\{X_{Test}\}^M$ consisting of M test images $X_{Test}=\{x_1,x_2,\ldots,x_M\} \in R^{dxM}$, where $x_m \in R^{dx(1\le m\le M)}$ is the mth sample of X_{Test}. Both the query image and test images in the repository are projected into the latent space via Ω_{Trained}. Then, a search is performed to identify the most similar embeddings in $E_{Test}=\{E_1,E_2,\ldots,E_M\} \in R^S$ to E_q as illustrated in Fig. 2.

4. Methods

4.1. Datasets

We have demonstrated the CBIR performance of *OCAM* on three distinct medical image datasets. The first dataset, KVASIR [32], contained endoscopic images from eight classes of gastrointestinal disease, with 1000 images per class. The second dataset, ISIC 2019 [33], contained dermoscopic images from eight classes of skin lesions, with the number of images in each class varying from 239 to 12875. The third dataset, curated X-RAY [34], contained radiographic images from four classes of lung disease, with the number of images in each class varying from 4497 to 5768. All images were downsampled to 256 × 256 pixels. 85% of the dataset was used for model training, and the remaining 15% was reserved for model testing.

4.2. Architectural details and model implementation

To capture the latent representations of medical images, we employed a Siamese neural network model with *OCAM* (Fig. 3). The weights of subnetworks responsible for processing anchor, positive and negative samples were tied. The subnetworks were designed by

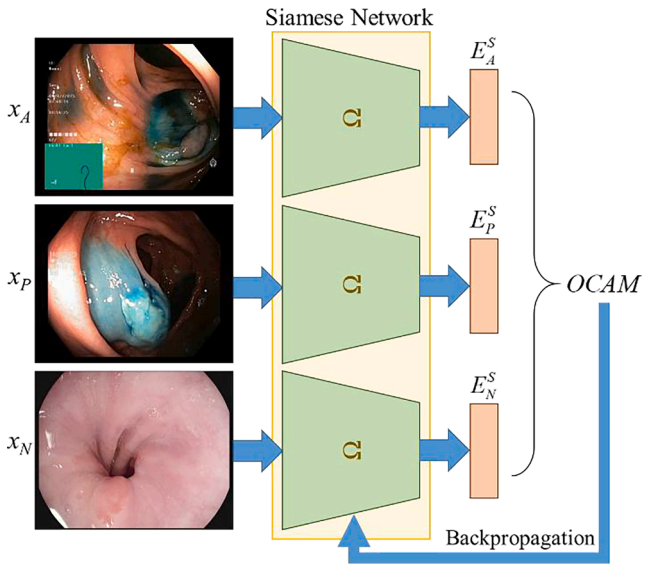

Fig. 3. Representational learning with *OCAM*. Embedding vectors E_A, E_P, and E_N for a randomly selected image triplet x_A, x_P, and x_N are generated with a Siamese network. The Siamese network contains backbone CNNs (Ω) with tied weights across anchor, positive and negative samples. CNN parameters are trained to minimize the *OCAM* loss.

considering popular convolutional neural network (CNN) architectures in computer vision: VGG16 [35], ResNet50 [36], InceptionV3 [37], MobileNetV2 [38], DenseNet169 [39], and EfficientNetB3 [40]. To obtain S-bit hash codes of image embeddings at the output of the Siamese network, the final fully connected (FC) layers in backbone CNNs were replaced with a dropout layer at 0.3 dropout rate and a dense layer of length S. Separate models were designed for $S = 16$ and $S = 64$. *OCAM* calculates the loss based on the similarity information using these embedding vectors of length S.

The models were implemented using the TensorFlow framework and executed on an NVidia RTX 3090 GPU. Model training was performed with the Adam optimizer, a batch size of 20 (corresponding to a selection of 60 images per batch due to triplet sampling), and a learning rate of 10^{-5}. Backbone CNNs were initialized with weights pretrained on ImageNet for object classification. The models were trained until convergence on each dataset, and because the dataset sizes varied, the number of epochs was 4500 for KVASIR, 45,000 for ISIC 2019, and 5000 for X-RAY datasets. To evaluate CBIR performance, query search with continuous codes in the Euclidean space of dense embedding vectors [1,7,8] and query search with binary codes in the Hamming space of the embedding vectors [26] were separately examined.

4.3. Competing methods

We present a comprehensive evaluation of *OCAM* by comparing it against state-of-the-art representational learning and classifier-guided methods. These comparisons are organized into two sections. For a focused assessment, the first section includes triplet-wise methods for representational learning based on the same framework as *OCAM* but implemented via recently proposed advanced triplet losses. The backbone architecture and training procedures for competing triplet-wise methods were identical to those for *OCAM* for a fair comparison. Meanwhile, for a more comprehensive assessment, the second section includes recent classifier-guided and representational-learning methods proposed for CBIR tasks in the literature. The network architecture and training procedures for this broader set of competing CBIR methods were adopted from the original studies. We adhered to the hyperparameters given in the original studies for each competing method described below.

4.3.1. Triplet-wise methods

Triplet: The *Triplet* method employs a traditional triplet loss function with α taken as 0.2 [8,11,14]:

$$Triplet(E_A, E_P, E_N) = \max(0, f(E_A, E_P) - f(E_A, E_N) + \alpha) \tag{7}$$

TriEP: The *TriEP* method uses weighting coefficients for distance measures in the triplet loss, and selects the hardest positive and negative samples [9].

$$TriEP(E_A, E_P, E_N) = [\sigma_1\sigma_2 \max(f(E_A, E_P)) - \beta_1\beta_2 \min(f(E_A, E_N)) + \alpha] \tag{8}$$

where σ_1, σ_2, β_1, β_2 represent the weighting coefficients, and $\alpha = 0.3$, $\sigma_1 = 2.04$, $\sigma_2 = 1.71$, $\beta_1 = 0.83$, $\beta_2 = 0.64$ were used.

WABT: The *WABT* method performs triplet-wise learning where the anchor sample is scaled prior to distance calculations to improve performance [10]:

$$WABT(E_A, E_P, E_N) = \max(0, f(rE_A, E_P) - f(rE_A, E_N) + \alpha) \tag{9}$$

where r denotes the scaling coefficient, and $\alpha = 1$ and $r = 3$ were used.

dmTri: The *dmTri* method uses a dynamic margin value by normalizing the loss function in the traditional triplet formulation with the sum of the *A-P* distance and α [41], with α taken as 0.2:

$$dmTri(E_A, E_P, E_N) = \max\left(0, 1 - \frac{f(E_A, E_N)}{f(E_A, E_P) + \alpha}\right) \tag{10}$$

CondTri: The *CondTri* method augments the traditional triplet loss with a weighted regularization term based on the individual distance measures [11], with $\alpha = 0.2$ and the regularization weight $\delta = 0.1$:

$$CondTri(E_A, E_P, E_N) = \max(0, f(E_A, E_P) - f(E_A, E_N) + \alpha) \; + \; \delta\left[\frac{f(E_A, E_P) + f(E_A, E_N)}{2}\right] \tag{11}$$

CTLL: The *CTLL* method augments the traditional triplet loss with a weighted and biased regularization term based on the difference norm of the embeddings for anchor and positive samples [12], with $\alpha = 1$, the regularization weight $\kappa = 0.01$, and the regularization bias $\gamma = 0.01$:

$$\begin{aligned} CTLL(E_A, E_P, E_N) = {} & \max(0, f(E_A, E_P) - f(E_A, E_N) + \alpha) \\ & + (\kappa(f(E_A - E_P)) - \gamma) \end{aligned} \tag{12}$$

4.3.2. Other Methods

Neural Codes: The *Neural Codes* method trains a backbone CNN architecture for a classification task and then discards its output layer to use it in a retrieval task [42]. VGG16 was taken as the backbone CNN, and the dimensionalities of two FC layers prior to softmax were changed to 2048 and S. In each dataset, the backbone CNN was trained to detect image classes by minimizing a point-wise categorical cross-entropy loss as in Eq. (7). For retrieval, the softmax layer was removed, and the image embedding was taken as

the vector of S-dimensional activations prior to the softmax layer in the trained CNN.

$$Neural\ Codes = -\frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{J} y_{k,j}\log\left(\Omega\left(x_{k,j};\theta\right)\right) \tag{13}$$

where K represents the total number of samples in the training repository, J represents the number of different classes in the training repository, and θ is the set of parameters of the Ω. A mini-batch size of 32 and a learning rate of 0.0001 were prescribed, and training was performed with the Adam optimizer.

Contrastive: The *Contrastive* method employs a traditional pair-wise contrastive loss on two anchor images (x_{A1}, x_{A2}) based on their similarity label (L, 0 = dissimilar, 1 = similar). For this purpose, two images are randomly selected from the repository: x_{A1}, x_{A2}. If these two anchor images belong to the same class, parameter L is set to 1, if they belong to different classes, parameter L is set to 0. As such, E_{A1} and E_{A2} can be derived by minimizing the following loss with α taken as 0.2 [7]:

$$Contrastive = L\frac{1}{2}f(E_{A1},E_{A2})^2 + (1-L)\frac{1}{2}\max(0,\alpha - f(E_{A1},E_{A2}))^2 \tag{14}$$

For *Contrastive*, a margin value of 0.2, a mini-batch size of 20 and a learning rate of 0.0001 were prescribed, and training was performed with the Adam optimizer.

Tr-Mixer-MLP: The vision transformer guided mixer MLP method (*Tr-Mixer-MLP*) employs a transformer architecture where the dense layers are replaced with a Mixer-MLP module [43]. It was trained for image classification, and the feature maps prior to the output layer were used for similarity measurements based on cosine distance. A mini-batch size of 32 and a learning rate of 0.001 were prescribed, and training was performed with the Adam optimizer.

QDDR: The query-driven distance recommendation (*QDDR*) approach employs ResNet50, a backbone CNN architecture for image classification, and utilizes the feature maps in the fully-connected layer before the output for similarity measurement based on cosine distance [20]. A mini-batch size of 32 and a learning rate of 0.01 were prescribed, and training was performed with an SGD optimizer with weight decay and momentum parameters set to 0.0001 and 0.99, respectively.

ABSM-Ret: The attention-based saliency map method (*ABSM-Ret*) generates saliency-weighted feature maps of images using an attention-guided CNN architecture [44]. Pair-wise representational learning is performed using a contrastive loss. A margin value of 1, a mini-batch size of 20 and a learning rate of 0.0001 were selected. Adam optimizer was used for training.

RIDH: The rotation-invariance deep hashing (*RIDH*) method uses a Cauchy rotation invariance loss for pair-wise representational learning between images at varying rotations [31]. A mini-batch size of 32, and a learning rate of 0.01 were used along with an SGD optimizer with weight decay and momentum parameters set to 0.0005 and 0.9, respectively.

X-MIR: The explainable medical image retrieval (*X-MIR*) method uses a CNN architecture for representational learning of image embeddings via the traditional triplet loss [45]. A constant margin value of 0.2, a mini-batch size of 16 and a learning rate of 0.0001 were prescribed, and training was performed with the Adam optimizer.

4.4. Performance metrics

CBIR performance was measured using the precision metric for a total of Z retrieved images (P@Z) and the mean average precision (mAP) metric [7,20,31]. Given a repository D with J image classes, P@Z was computed as the across-class average of class-specific (P@Z)$_j$. To compute (P@Z)$_j$, a single test image from the jth class was taken as the query, and retrieval was attempted on the entire test set, excluding the query image. This process was repeated across all possible query images for the jth class. Afterwards, class-specific performance (P@Z)$_j$ and P@Z were computed:

$$\begin{aligned}(P@Z)_j &= \frac{1}{n_j Z}\sum_{i=1}^{n_j}\sum_{z=1}^{Z} t\left(x_i^j, NN\left(x_i^j,Z\right)_z\right)\\ P@Z &= \frac{1}{J}\sum_{j=1}^{J}(P@Z)_j\end{aligned} \tag{15}$$

where n_j denotes the number of samples in the jth class, x_i^j is the ith query image from the jth class, $NN(.)_z$ represents the zth element retrieved from the repository among the set of Z images, and t is an indicator function that returns 1 if its inputs are from the same class, and returns 0 otherwise. (mAP)$_j$ and mAP were calculated as follows:

$$\begin{aligned}(mAP)_j &= \frac{1}{n_j Z}\sum_{i=1}^{n_j}\sum_{z=1}^{Z}\frac{t\left(x_i^j, NN\left(x_i^j,Z\right)_z\right)}{z}\\ mAP &= \frac{1}{j}\sum_{j=1}^{J}(mAP)_j\end{aligned} \tag{16}$$

Note that a high P@Z score can be achieved by retrieving a large number of images from the same class as the query image. Meanwhile, mAP does not only consider the raw number of correctly retrieved images but it also reflects the order in which images are

retrieved. In particular, a high mAP score can be attained when images from the same class as the query are retrieved in higher ranks among Z images compared to images from other classes.

5. Results

5.1. Ablation studies

Several ablation studies were performed to demonstrate the value of the individual components in *OCAM*, including the backbone CNN, the *P-N* distance loss component, and the adaptive margin value. First, retrieval performance was evaluated for *OCAM* variants based on six different backbone CNNs, and the resulting mAP performances with $S = 64$ are given in Table 1 (E represents Embedding space, and H represents Hamming space). Across all CBIR tasks, the variants with VGG16 produced the most optimal or near-optimal results, thus it was chosen as the backbone CNN in the remaining experiments.

The retrieval performance of *OCAM* was also evaluated comparatively against a variant ablated of the *P-N* distance $f(E_P,E_N)$ (w/o $f(E_P,E_N)$), a variant ablated of the adaptive margin $\alpha_{adaptive}$ (w/o $\alpha_{adaptive}$), and a variant ablated of both the *P-N* distance and the adaptive margin (w/o $f(E_P,E_N)$ and $\alpha_{adaptive}$). Performances of variant models are shown in Fig. 4 with $S = 64$. Overall, *OCAM* outperforms all ablated variants and the performance benefits grow towards larger Z values. Furthermore, both $f(E_P,E_N)$ and $\alpha_{adaptive}$ elements in *OCAM* contribute to CBIR performance in terms of P@Z and mAP. Both Euclidean and Hamming spaces exhibit nearly similar observations, however, it is anticipated that there will be a decrease in performance for all methods when operating in the Hamming space due to the binarization of embeddings.

The primary motivation of *OCAM* is to enhance the discernibility of latent space representations for images from separate classes. We hypothesized that if *OCAM* improves representational discernibility over traditional triplet loss, then the learned embedding vectors should be better segregated in the embedding space. To validate this hypothesis, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) [8] to project the embedding vectors obtained using OCAM and Triplet into a two-dimensional space. The image projections in the t-SNE space are displayed in Fig. 5. It can be observed that for *OCAM*, images from different classes project to spatially segregated clusters that are well separated from each other. In contrast, the separation between clusters is relatively lower for Triplet, and samples from distinct clusters can occur in spatially proximate locations. These findings indicate that *OCAM* improves inter-class discernibility over the traditional triplet method.

To assess the influence of this representational discernibility on retrieved images, we visually inspected images retrieved based on *OCAM* and *Triplet*. Fig. 6 depicts $Z = 10$ images retrieved in response to a query image randomly selected from the test set of each of the three datasets. Overall, the images retrieved using *OCAM* were more visually similar to the query image. It's worth noting that with CBIR methods, there may be instances where images from different categories are retrieved. An essential performance criterion for CBIR methods is the retrieve rank of opponent-class samples (which is also captured in the mAP metric). The retrieve rank of opponent samples in *OCAM* is lower than that for *Triplet*, indicating that *OCAM* is more resilient against erroneous sample selection.

5.2. Performance comparisons against Triplet-Wise methods

This section presents a comparative evaluation of OCAM for CBIR tasks against state-of-the-art triplet losses based representational-learning methods. Competing methods included traditional triplet learning (*Triplet*) [8], expansion-pool tri-hard learning (*TriEP*) [9], weighted anchor based triplet learning (*WABT*) [10], dynamic margin triplet learning (*dmTri*) [41], conditional triplet learning (*CondTri*) [11], and constrained triplet loss layer learning (*CTLL*) [12]. Experiments were conducted in the continuous Euclidean space to demonstrate the full performance of learned representations, and in binarized Hamming space to demonstrate the performance in real-time applications that require computational efficiency.

Euclidean Space: Tables 2-4 list retrieval performances of triplet-wise competing methods on the test sets of the KVASIR, ISIC 2019, and X-RAY datasets, respectively. Note that the number of samples included in the minority class varied across datasets, which places a practical limit on the number of images that can be retrieved. Accordingly, retrieval was performed for up to $Z = 150$ images in KVASIR and X-RAY, and up to $Z = 35$ images in ISIC 2019 that had a more dramatic class imbalance. Compared to the second-best method, *OCAM* improves mAP by 1.57% at $S = 16$ and 1.44% at $S = 64$ in KVASIR, by 2.03% at $S = 16$ and 3.80% at $S = 64$ in ISIC 2019, and by 0.25% at $S = 16$ and 2.21% at $S = 64$ in X-RAY.

Hamming Space: Tables 5-7 list retrieval performances of triplet-wise competing methods on the test sets of the KVASIR, ISIC 2019,

Table 1
Retrieval Performance of *OCAM* Across Backbone CNNs.

	KVASIR ($S = 64$)		ISIC 2019 ($S = 64$)		X-RAY ($S = 64$)	
	E	H	E	H	E	H
VGG16 [35]	**88.74**	**85.93**	72.95	**71.48**	**87.32**	**86.41**
ResNet50 [36]	87.03	82.22	72.44	70.91	86.52	85.71
InceptionV3 [37]	84.49	80.13	70.23	69.16	84.60	83.71
MobileNetV2 [38]	82.13	75.62	72.16	70.76	83.57	82.66
DenseNet169 [39]	88.52	84.37	72.98	71.20	85.98	85.34
EfficientNetB3 [40]	85.47	81.33	**73.13**	71.13	83.76	82.52

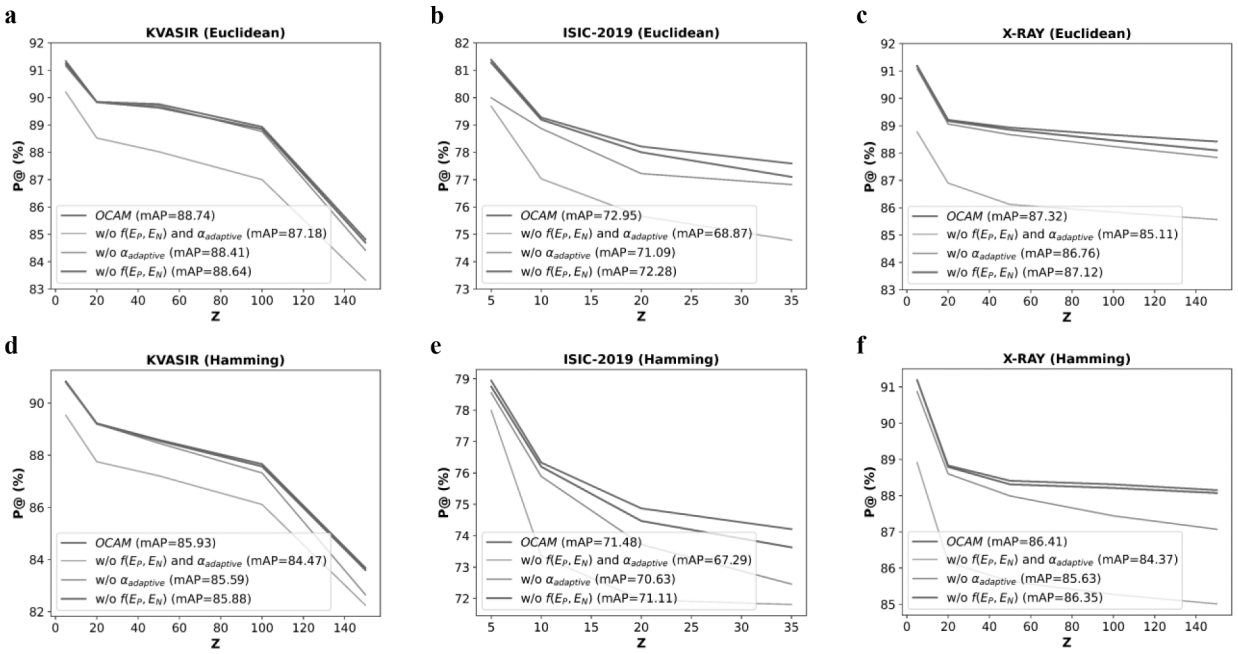

Fig. 4. Precision performances of *OCAM* variants were measured based on 64-dimensional embedding vectors. P@Z metrics are plotted for the KVASIR (a,d), ISIC 2019 (b,e), and X-RAY (c,f) datasets across various numbers of retrieved images (*Z*). Results for Euclidean space are given in the top row, and those for Hamming space are given in the bottom row. mAP metrics for each variant are listed in parentheses.

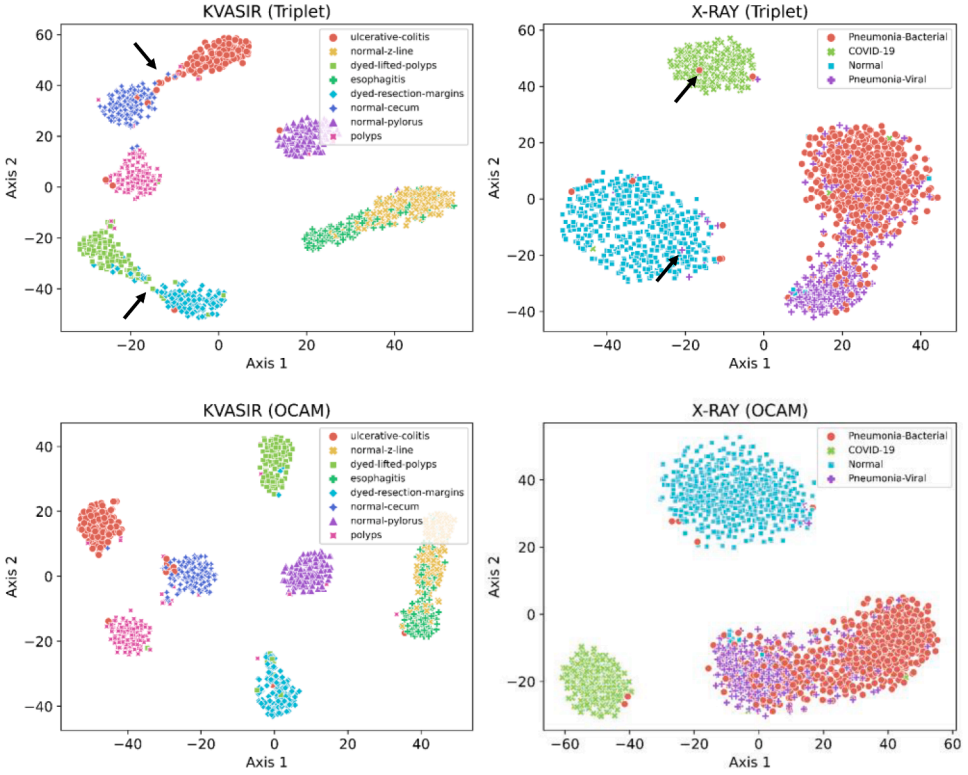

Fig. 5. The latent space distribution of images was visualized using two-dimensional t-SNE projections of learned embedding vectors. Representative results are displayed for embeddings learned based on the traditional triplet method (*Triplet*) versus *OCAM*. *OCAM* yields more discrete clusters than *Triplet*. Black arrows highlight problematic samples with insufficient inter-class segregation. The distinct classes and their labels are described in respective legends for each dataset.

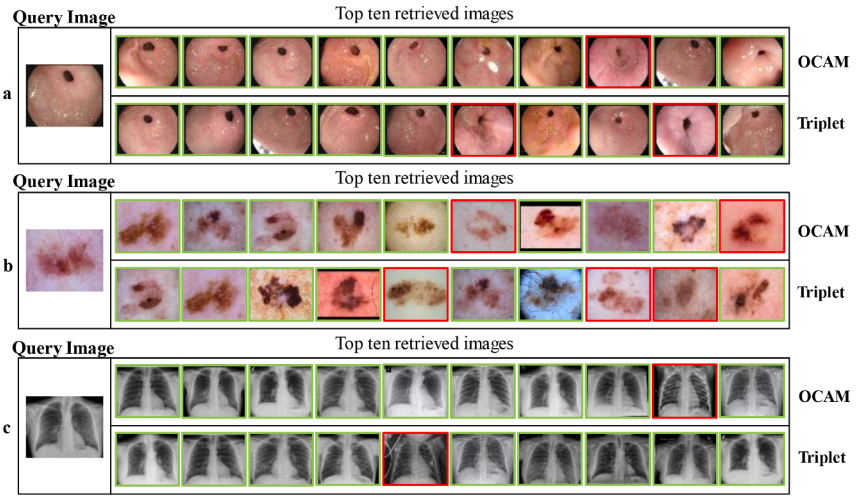

Fig. 6. Representative retrieval results for $Z = 10$ images given a single query image. Results are shown for the a) KVASIR, b) ISIC-2019, c) X-RAY datasets. Images retrieved from the same class as the query image are marked in green bounding boxes, whereas opponent-class samples are marked in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and X-RAY datasets, respectively. As expected, the performances of these methods are moderately lower in the Hamming versus Euclidean space due to the loss of information during the binarization of embedding vectors. In general, *OCAM* outperforms competing methods in CBIR performance. Compared to the second-best method, *OCAM* improves mAP by 3.85% at $S = 16$ and 1.46% at $S = 64$ in KVASIR, by 1.07% at $S = 16$ and 2.71% at $S = 64$ in ISIC 2019, and by 2.01% at $S = 64$ in X-RAY. Meanwhile, it yields a moderately lower mAP by 0.22% compared to *WABT* with Hamming codes at $S = 16$ in X-RAY.

Overall, we observe that the performance benefits of *OCAM* against competing triplet-wise methods are relatively larger in the ISIC 2019 dataset versus the KVASIR and XRAY datasets, where performance metrics are generally higher. It is important to note that the difficulties of the representational learning task and subsequent image retrieval task grow with higher class imbalances in the repository. Thus, this finding suggests that *OCAM* shows improved reliability against class imbalances compared to baseline methods.

5.3. Performance comparisons against other CBIR methods

Next, *OCAM* was comparatively demonstrated for CBIR tasks against other SOTA image retrieval approaches. Competing methods included a classifier-guided model (*Neural Codes*) [42], a pair-wise representational learning model (*Contrastive*) [7], a classifier-guided vision transformer and mixer-MLP model (*Tr-Mixer-MLP*) [43], a classifier-guided query-driven distance recommendation (*QDDR*) technique [20], a pair-wise representational learning model based via attention-based saliency map guided CNN (*ABSM-Ret*) [44], a classifier-guided model based on rotation-invariance deep hashing (*RIDH*) [31], a triplet-wise representational learning model with saliency maps for explainable medical image retrieval (*X-MIR*) [45]. Experiments were conducted in Euclidean space and binarized Hamming space for $S = 16$ and 64. Table 8 lists the mAP performances of competing methods on the test sets of the KVASIR, ISIC 2019, and X-RAY datasets. In Euclidean space, we find that *OCAM* outperforms the second-best competing method in mAP performance by 0.27% at $S = 16$ and 0.37% at $S = 64$ for the KVASIR dataset, by 3.36% at $S = 16$ and 2.72% at $S = 64$ for the ISIC 2019 dataset, and by 0.56% at $S = 16$ and 1.25% at $S = 64$ for the X-RAY dataset. In Hamming space, we find that *OCAM* outperforms the second-best competing method in mAP performance by 0.12% at $S = 16$ and 0.65% at $S = 64$ for the KVASIR dataset, by 2.79% at $S = 16$ and 3.52% at $S = 64$ for the ISIC 2019 dataset, and by 0.67% at $S = 16$ and 0.56% at $S = 64$ for the X-RAY dataset. Overall, the performance benefits of *OCAM* over competing methods are relatively more pronounced in the ISIC 2019 dataset, which manifests a

Table 2
Retrieval Performance ın KVASIR Dataset (Euclidean).

	$S = 16$						$S = 64$					
	P@5	P@20	P@50	P@100	P@150	mAP	P@5	P@20	P@50	P@100	P@150	mAP
Triplet [8]	90.13	87.61	86.88	85.52	81.32	85.22	90.20	88.52	88.02	87.00	83.33	87.18
TriEP [9]	87.83	85.59	84.69	83.06	78.64	83.01	86.78	83.54	81.98	79.83	75.97	80.10
WABT [10]	81.45	76.67	75.23	74.26	72.17	74.80	82.45	78.28	76.32	74.96	72.40	75.71
dmTri [41]	90.47	88.35	87.38	85.97	81.90	85.73	90.32	88.40	87.56	86.16	82.01	86.27
CondTri [11]	87.48	85.58	84.81	83.26	79.75	83.63	90.96	88.84	88.13	87.22	83.77	87.30
CTLL [12]	89.72	87.07	86.47	85.24	80.89	85.09	89.88	87.89	87.39	86.68	82.87	86.71
OCAM	90.75	88.97	88.64	87.40	84.13	**87.30**	91.33	89.84	89.76	88.93	84.82	**88.74**

Table 3
Retrieval Performance ın ISIC 2019 Dataset (Euclidean).

	$S = 16$					$S = 64$				
	P@5	P@10	P@20	P@35	mAP	P@5	P@10	P@20	P@35	mAP
Triplet [8]	77.70	74.84	73.21	72.48	67.92	79.68	77.03	75.66	74.79	68.87
TriEP [9]	76.70	73.55	72.28	71.33	66.45	78.60	75.26	73.61	72.49	67.84
WABT [10]	73.44	70.35	68.76	67.78	63.64	76.68	73.59	71.99	71.10	66.67
dmTri [41]	77.75	74.81	73.25	72.39	68.03	80.26	78.09	76.92	76.15	71.27
CondTri [11]	78.08	75.16	73.98	72.94	68.74	79.71	77.24	76.22	75.48	70.75
CTLL [12]	77.75	74.75	73.19	72.21	67.50	78.64	75.74	74.32	71.18	69.15
OCAM	79.29	77.00	75.79	75.20	**70.77**	81.38	79.27	78.21	77.59	**72.95**

Table 4
Retrieval Performance ın X-RAY Dataset (Euclidean).

	$S = 16$						$S = 64$					
	P@5	P@20	P@50	P@100	P@150	mAP	P@5	P@20	P@50	P@100	P@150	mAP
Triplet [8]	88.60	86.13	85.53	85.41	85.21	84.48	88.77	86.90	86.12	85.84	85.57	85.11
TriEP [9]	88.10	85.45	84.82	84.77	84.63	84.06	88.32	86.20	85.70	85.42	85.18	84.68
WABT [10]	89.28	87.39	86.85	86.58	86.31	85.41	87.47	85.40	85.22	85.18	85.02	84.68
dmTri [41]	88.65	86.03	85.37	85.09	84.94	84.12	88.36	86.14	85.78	85.64	85.40	84.82
CondTri [11]	88.78	86.63	86.02	85.67	85.48	84.63	88.74	86.45	85.82	85.41	85.16	84.63
CTLL [12]	88.32	85.78	85.07	84.84	84.69	84.07	88.53	86.25	85.96	85.72	85.47	84.88
OCAM	90.42	88.49	87.84	87.53	87.20	**85.66**	91.19	89.21	88.93	88.66	88.42	**87.32**

Table 5
Retrieval Performance ın KVASIR Dataset (Hamming).

	$S = 16$						$S = 64$					
	P@5	P@20	P@50	P@100	P@150	mAP	P@5	P@20	P@50	P@100	P@150	mAP
Triplet [8]	86.42	83.68	81.84	79.35	73.47	75.07	89.53	87.75	87.21	86.11	82.25	84.47
TriEP [9]	84.10	82.93	80.85	78.38	72.99	72.74	83.93	81.06	79.78	78.13	74.57	71.03
WABT [10]	76.08	74.66	73.62	73.10	70.67	63.94	74.83	73.96	73.45	73.23	71.20	67.13
dmTri [41]	87.81	85.38	84.50	82.26	76.80	79.37	88.93	87.08	86.16	84.81	80.01	83.97
CondTri [11]	84.73	83.25	82.02	79.61	74.04	72.43	88.98	87.25	86.38	84.87	80.90	82.73
CTLL [12]	84.92	82.56	80.53	77.44	71.94	74.42	86.62	87.45	86.81	85.87	81.78	84.38
OCAM	89.10	87.28	86.33	84.95	79.32	**83.22**	90.83	89.22	88.59	87.66	83.68	**85.93**

Table 6
Retrieval Performance ın ISIC 2019 Dataset (Hamming).

	$S = 16$					$S = 64$				
	P@5	P@10	P@20	P@35	mAP	P@5	P@10	P@20	P@35	mAP
Triplet [8]	74.74	72.57	70.88	69.97	64.15	77.99	73.36	71.96	71.91	67.29
TriEP [9]	71.50	71.15	69.21	68.16	62.12	75.72	73.03	71.22	70.08	64.14
WABT [10]	61.57	66.12	66.91	65.58	61.60	61.82	62.51	64.96	64.17	59.73
dmTri [41]	72.95	70.60	69.55	68.70	62.31	77.91	75.34	73.91	73.02	68.44
CondTri [11]	74.91	72.88	71.01	70.11	64.91	78.22	75.45	74.00	73.14	68.77
CTLL [12]	72.59	71.48	70.52	69.59	64.73	77.47	74.73	73.05	71.97	67.94
OCAM	75.63	73.93	72.12	71.36	**65.98**	78.94	76.33	74.87	74.21	**71.48**

higher degree of imbalanced compared to the KVASIR and X-RAY datasets. This suggests that *OCAM* is capable of generating robust embedding vectors even on more imbalanced datasets.

6. Dıscussıon

CBIR methods aim to provide fast and automated access to visually similar images from a medical repository, enabling downstream assessments with visual examples. Current CBIR approaches typically rely on the learning of embedding vectors with high representation capability for visual image features, as subsequent retrieval performance depends critically on the representational quality of the embedding vectors [9,20,31,45]. In this study, we introduced a novel triplet-wise representational learning method, *OCAM*, for improved CBIR performance on medical image repositories. *OCAM* leverages triplet learning with an improved objective that considers distances between positive and negative classes and an adaptive margin value. Explicit consideration of $f(P,N)$ enables *OCAM* to improve inter-class segregation in the embedding space, while the adaptive margin value improves performance by automatically

Table 7
Retrieval Performance ın X-RAY Dataset (Hamming).

	$S = 16$						$S = 64$					
	P@5	P@20	P@50	P@100	P@150	mAP	P@5	P@20	P@50	P@100	P@150	mAP
Triplet [8]	85.85	83.79	83.43	82.96	82.67	81.29	88.91	86.16	85.63	85.27	85.01	84.37
TriEP [9]	84.41	83.35	83.05	82.44	81.95	80.50	87.41	85.32	85.06	84.84	84.65	84.09
WABT [10]	87.28	85.89	85.70	85.30	84.82	**83.11**	87.89	85.58	85.22	85.01	84.89	84.40
dmTri [41]	86.10	83.93	82.97	82.45	82.22	80.43	88.55	86.11	85.49	85.21	85.00	83.90
CondTri [11]	85.76	83.99	83.45	82.81	82.28	79.55	88.01	85.54	84.91	84.62	84.40	83.38
CTLL [12]	86.17	84.12	83.44	82.88	82.41	80.98	88.64	86.19	85.67	85.39	85.06	84.16
OCAM	88.83	86.60	86.02	85.55	85.25	82.89	91.19	88.83	88.41	88.31	88.15	**86.41**

Table 8
Evaluation of Performance in Relation to Other CBIR Methods.

	Euclidean						Hamming					
	KVASIR		ISIC 2019		X-RAY		KVASIR		ISIC 2019		X-RAY	
	S = 16	S = 64	S = 16	S = 64	S = 16	S = 64	S = 16	S = 64	S = 16	S = 64	S = 16	S = 64
Neural Codes [42]	78.07	79.36	56.15	57.90	81.17	81.86	75.17	78.89	55.57	56.38	79.07	79.25
Contrastive [7]	80.63	85.59	58.85	60.27	84.40	84.72	72.96	81.23	54.74	58.11	76.44	84.18
Tr-Mixer-MLP [43]	83.15	85.66	59.10	60.87	82.88	85.28	77.91	81.51	56.81	59.16	79.16	83.87
QDDR [20]	81.74	83.17	59.62	62.22	81.73	83.11	79.26	81.11	56.17	59.43	81.10	82.73
ABSM-Ret [44]	83.11	87.24	60.44	64.09	83.86	85.87	81.59	83.77	56.48	61.24	81.23	84.96
RIDH [31]	87.03	88.37	65.29	68.26	85.10	85.93	83.10	85.28	61.02	64.07	81.87	85.85
X-MIR [45]	86.81	88.30	67.41	70.23	84.95	86.07	82.95	85.02	63.19	67.96	82.22	85.27
OCAM	**87.30**	**88.74**	**70.77**	**72.95**	**85.66**	**87.32**	**83.22**	**85.93**	**65.98**	**71.48**	**82.89**	**86.41**

tuning the margin value depending on the learned representations at each stage of the training process. Our results clearly indicate that *OCAM* outperforms a state-of-the-art point-wise method (*Neural Codes*) by 9.15%, pair-wise methods (*Contrastive* and *ABSM-Ret*) by 5.90%, and triplet-wise methods (*X-MIR*) by %1.52 on average.

In our study, we investigated how the code length and type affect CBIR performance, which is crucial for image similarity assessments. We conducted CBIR tasks with two code lengths, $S = 16$ and 64, and observed that retrieval performance improved as S increased. On average, the increase in mAP performance of *OCAM* is 2.9% when the S is elevated from 16 to 64. We also assessed CBIR tasks based on continuous Euclidean and binary Hamming codes. We observed that mAP performance in Hamming space is lower than in the Euclidean space due to information loss during the binarization process. On average, the increase in mAP performance of *OCAM* in Euclidean space is 2.8% compared to the Hamming space. Lastly, CIBR performance can also depend on the native imbalance between different classes of images in the repository. A general inspection of the results in Tables 2-8 suggests that CBIR performance is higher in relatively more balanced KVASIR and X-RAY datasets than in the imbalanced ISIC 2019 dataset.

Several lines of improvement might enable *OCAM* to further its performance in CBIR tasks. Following the general trend in retrieval performance going from point-wise to triplet-wise methods, the quality of the learned embedding vectors may be further improved by adopting advanced loss functions based on quadruplet learning that utilizes a larger number of image samples [46]. Architectures that explicitly leverage self-attention mechanisms or vision transformers might enable the capture of more representative embedding vectors for images by better modeling spatial context [47]. A task-agnostic approach to more sensitively capture the distributional properties of medical images might employ recent diffusion models [48,49]. By using joint embedding of text and image data where both a radiological report and a corresponding image are available, semantic similarities among medical images might be more effectively captured [50]. Lastly, it may be possible to perform unsupervised retrieval with *OCAM* by learning the latent representations of images using unconditional generative adversarial network models in cases where class information is unavailable.

7. Conclusion

Triplet-wise methods for learning image representations promise superior CBIR performance over both classifier-guided models and representational learning with point- or pair-wise methods. However, the traditional triplet formulation can suffer from suboptimal segregation between positive and negative samples, and requires manual tuning of a margin value. Here, we introduced a new triplet-wise method, *OCAM*, that addresses these limitations for improved CBIR performance. *OCAM* was demonstrated on three medical datasets from divergent domains for CBIR tasks executed in Euclidean and Hamming spaces. Our results indicate that *OCAM* outperforms competing state-of-the-art CBIR methods based on classifier-guided and representational-learning frameworks, including several existing triplet-wise methods. Elevated performance is observed over competing methods particularly for imbalanced datasets. As such, *OCAM* holds great promise for automating query search in CBIR systems that aim to improve diagnostic accuracy while minimizing processing delays. In future studies, we aim to investigate potential enhancements in CBIR performance by utilizing advanced loss functions that leverage a larger number of image samples, by incorporating vision transformer architectures that can

improve the capture of long-range spatial context in medical images, and by employing generative modeling frameworks for more sensitive and possibly unsupervised learning of latent image representations. For applications where pairs of medical images and respective radiology reports are available, we will further investigate the effects of utilizing multi-modal embedding vectors of image-text data on retrieval performance.

CRediT authorship contribution statement

Şaban Öztürk: Conceptualization, Methodology, Investigation, Software, Writing – original draft, Visualization. **Emin Çelik:** Conceptualization, Methodology, Validation, Writing – original draft, Visualization. **Tolga Çukur:** Conceptualization, Methodology, Writing – review & editing, Formal analysis, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] M. Garg, G. Dhiman, A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants, Neural Computing and Applications 33 (2021) 1311–1328.
[2] S. Chavda, M. Goyani, Hybrid Approach to Content-Based Image Retrieval Using Modified Multi-Scale LBP and Color Features, SN Computer Science 1 (2020) 305.
[3] X. Zhang, C. Bai, K. Kpalma, OMCBIR: Offline mobile content-based image retrieval with lightweight CNN optimization, Displays 76 (2023), 102355.
[4] A. Alzu'bi, A. Amira, N. Ramzan, Compact Root Bilinear CNNs for Content-Based Image Retrieval, 2016 International Conference on Image, Vision and Computing (Icivc 2016), (2016) 41-45.
[5] Ş. Öztürk, Hash code generation using deep feature selection guided siamese network for content-based medical image retrieval, Gazi University Journal of, Science 34 (2021) 733–746.
[6] D. Pathak, U.S.N. Raju, Content-based image retrieval for super-resolutioned images using feature fusion: Deep learning and hand crafted, Concurrency and Computation: Practice and Experience (2022) e6851.
[7] S. Deepak, P. Ameer, Retrieval of brain MRI with tumor using contrastive loss based similarity on GoogLeNet encodings, Computers in Biology and Medicine 125 (2020), 103993.
[8] N. Loiseau-Witon, R. Kéchichian, S. Valette, A. Bartoli, Learning 3D medical image keypoint descriptors with the triplet loss, International Journal of Computer Assisted Radiology and Surgery 17 (2022) 141–146.
[9] Z. Yu, W. Qin, L. Tahsin, Z. Huang, TriEP: Expansion-Pool TriHard Loss for Person Re-Identification, Neural Processing Letters 54 (2022) 2413–2432.
[10] T. Bui, L. Ribeiro, M. Ponti, J. Collomosse, Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network, Computer Vision and Image Understanding 164 (2017) 27–37.
[11] D. Shi, M. Orouskhani, Y. Orouskhani, A conditional Triplet loss for few-shot learning and its application to image co-segmentation, Neural Networks 137 (2021) 54–62.
[12] H.W.F. Yeung, J. Li, Y.Y. Chung, Improved performance of face recognition using CNN with constrained triplet loss layer, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 1948–1955.
[13] D. Cheng, Y. Gong, W. Shi, S. Zhang, Person re-identification by the asymmetric triplet and identification loss function, Multimedia Tools and Applications 77 (2018) 3533–3550.
[14] Ş. Öztürk, T. Çukur, Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets, IEEE Journal of Biomedical and Health Informatics 26 (2022) 4679–4690.
[15] A. Latif, A. Rasheed, U. Sajid, J. Ahmed, N. Ali, N.I. Ratyal, B. Zafar, S.H. Dar, M. Sajid, T. Khalil, Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review, Mathematical Problems in Engineering 2019 (2019) 1–21.
[16] S.R. Dubey, A Decade Survey of Content Based Image Retrieval Using Deep Learning, IEEE Transactions on Circuits and Systems for Video Technology 32 (2022) 2687–2704.
[17] J. Pradhan, A.K. Pal, H. Banka, P. Dansena, Fusion of region based extracted features for instance- and class-based CBIR applications, Applied Soft Computing 102 (2021), 107063.
[18] A. Ponomarev, H.S. Nalamwar, I. Babakov, C.S. Parkhi, G. Buddhawar, Content-Based Image Retrieval Using Color, Texture and Shape Features, Key Engineering Materials 685 (2016) 872–876.
[19] E. Rosten, R. Porter, T. Drummond, Faster and Better: A Machine Learning Approach to Corner Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 105–119.
[20] W. Barhoumi, A. Khelifa, Skin lesion image retrieval using transfer learning-based approach for query-driven distance recommendation, Computers in Biology and Medicine 137 (2021), 104825.
[21] Y.-H.-H. Tsai, H. Zhao, M. Yamada, L.-P. Morency, R.R. Salakhutdinov, Neural methods for point-wise dependency estimation, Advances in Neural Information Processing Systems 33 (2020) 62–72.
[22] Ş. Öztürk, Class-driven content-based medical image retrieval using hash codes of deep features, Biomedical Signal Processing and Control 68 (2021), 102601.
[23] J.Y.H. Ng, F. Yang, L.S. Davis, Exploiting Local Features from Deep Networks for Image Retrieval, Ieee Comput Soc Conf (2015) 53–61.
[24] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, in, 2015, pp. arXiv:1511.05879.
[25] S.S. Husain, M. Bober, REMAP: Multi-Layer Entropy-Guided Pooling of Dense CNN Features for Image Retrieval, IEEE Transactions on Image Processing 28 (2019) 5201–5213.
[26] Z. Zhang, Q. Zou, Y. Lin, L. Chen, S. Wang, Improved deep hashing with soft pairwise similarity for multi-label image retrieval, IEEE Transactions on Multimedia 22 (2019) 540–553.
[27] M.L. Ha, V. Blanz, Deep ranking with adaptive margin triplet loss, arXiv preprint arXiv:2107.06187, (2021).
[28] X.A. Zhao, H. Qi, R. Luo, L. Davis, A weakly supervised adaptive triplet loss for deep metric learning, in: 2019 Ieee/Cvf International Conference on Computer Vision Workshops (Iccvw), 2019, pp. 3177–3180.

[29] W. Xie, H. Wu, Y. Tian, M. Bai, L. Shen, Triplet Loss With Multistage Outlier Suppression and Class-Pair Margins for Facial Expression Recognition, IEEE Transactions on Circuits and Systems for Video Technology 32 (2022) 690–703.
[30] X. Luo, H. Wang, D. Wu, C. Chen, M. Deng, J. Huang, X.-S. Hua, A survey on deep hashing methods, ACM Transactions on Knowledge Discovery from Data 17 (1) (2023) 1–50.
[31] Y. Zhang, F. Xie, X. Song, Y. Zheng, J. Liu, J. Wang, Dermoscopic image retrieval based on rotation-invariance deep hashing, Medical Image Analysis 77 (2022), 102301.
[32] K. Pogorelov, K.R. Randel, C. Griwodz, S.L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P.T. Schmidt, Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, in: Proceedings of the 8th ACM on Multimedia Systems Conference, 2017, pp. 164-169.
[33] N.C. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, 2018, pp. 168–172.
[34] U. Sait, K. Lal, S. Prajapati, R. Bhaumik, T. Kumar, S. Sanjana, K. Bhalla, Curated dataset for COVID-19 posterior-anterior chest radiography images (X-Rays), Mendeley Data 1 (2020).
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, (2014).
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.
[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.
[38] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861, (2017).
[39] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700-4708.
[40] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105-6114.
[41] S. Zakharov, W. Kehl, B. Planche, A. Hutter, S. Ilic, 3D Object Instance Recognition and Pose Estimation Using Triplet Loss with Dynamic Margin, Ieee Int C Int Robot (2017) 552–559.
[42] A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky, Neural codes for image retrieval, in: European conference on computer vision, Springer, 2014, pp. 584–599.
[43] Q.-H. Trinh, M.-V. Nguyen, Endoscopy Image Retrieval by Mixer Multi-Layer Perceptron, 16th Conference on Computer Science and Intelligence Systems, 26 (2021) 223-226.
[44] Ş. Öztürk, A. Alhudhaif, K. Polat, Attention-based end-to-end CNN framework for content-based X-ray image retrieval, Turkish Journal of Electrical Engineering & Computer Sciences 29 (2021) 2680–2693.
[45] B. Hu, B. Vasu, A. Hoogs, X-MIR: EXplainable Medical Image Retrieval, Ieee Wint Conf Appl (2022) 1544–1554.
[46] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 403-412.
[47] O. Dalmaz, M. Yurt, T. Çukur, ResViT: residual vision transformers for multimodal medical image synthesis, IEEE Transactions on Medical Imaging 41 (2022) 2598–2614.
[48] A. Güngör, S.U. Dar, Ş. Öztürk, Y. Korkmaz, G. Elmas, M. Özbey, T. Çukur, Adaptive Diffusion Priors for Accelerated MRI Reconstruction, in, 2022, pp. arXiv: 2207.05876.
[49] M. Özbey, O. Dalmaz, S.U. Dar, H.A. Bedel, Ş. Özturk, A. Güngör, T. Çukur, Unsupervised Medical Image Translation with Adversarial Diffusion Models, in, 2022, pp. arXiv:2207.08208.
[50] N. Malali, Y. Keller, Learning to Embed Semantic Similarity for Joint Image-Text Retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2022) 10252–10260.