Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064403

ARTICLE

Efficient Method for Trademark Image Retrieval: Leveraging Siamese and Triplet Networks with Examination-Informed Loss Adjustment

Thanh Bui-Minh[1], Nguyen Long Giang[1] and Luan Thanh Le[2,*]

[1]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, 100000, Vietnam
[2]Department of Business, Greenwich Vietnam, FPT University, Hanoi, 100000, Vietnam

*Corresponding Author: Luan Thanh Le. Email: luanlt13@fe.edu.vn

Received: 14 February 2025; Accepted: 22 April 2025; Published: 09 June 2025

ABSTRACT: Image-based similar trademark retrieval is a time-consuming and labor-intensive task in the trademark examination process. This paper aims to support trademark examiners by training Deep Convolutional Neural Network (DCNN) models for effective Trademark Image Retrieval (TIR). To achieve this goal, we first develop a novel labeling method that automatically generates hundreds of thousands of labeled similar and dissimilar trademark image pairs using accompanying data fields such as citation lists, Vienna classification (VC) codes, and trademark ownership information. This approach eliminates the need for manual labeling and provides a large-scale dataset suitable for training deep learning models. We then train DCNN models based on Siamese and Triplet architectures, evaluating various feature extractors to determine the most effective configuration. Furthermore, we present an Adapted Contrastive Loss Function (ACLF) for the trademark retrieval task, specifically engineered to mitigate the influence of noisy labels found in automatically created datasets. Experimental results indicate that our proposed model (EfficientNet_v21_Siamese) performs best at both True Negative Rate (TNR) threshold levels, TNR = 0.9 and TNR = 0.95, with respective True Positive Rates (TPRs) of 77.7% and 70.8% and accuracies of 83.9% and 80.4%. Additionally, when testing on the public trademark dataset METU_v2, our model achieves a normalized average rank (NAR) of 0.0169, outperforming the current state-of-the-art (SOTA) model. Based on these findings, we estimate that considering only approximately 10% of the returned trademarks would be sufficient, significantly reducing the review time. Therefore, the paper highlights the potential of utilizing national trademark data to enhance the accuracy and efficiency of trademark retrieval systems, ultimately supporting trademark examiners in their evaluation tasks.

KEYWORDS: Trademark; image retrieval; similar search; similar retrieval; content-based image retrieval; similar ranking; contrastive learning; Siamese; triplet; citation list

1 Introduction

In the globalized era, trademark registration has become a crucial element in safeguarding IP rights and preventing legal disputes for companies worldwide [1]. In developing countries, trademark image verification predominantly relies on conventional methodologies, where examiners manually categorize trademark images into classification tables [2]. This process is not only time-consuming but also prone to errors, especially as the number of registered trademarks in the database grows over time. At Intellectual Property (IP) Offices, examiners analyze the visual characteristics of trademark images and classify them using the Vienna Classification (VC) international design codes. However, the classification of images based on the fixed VC has several limitations. First, the manual VC performed by examiners relies heavily on their subjective judgment, and different examiners may assign different codes to the same image. This leads

to many cases of similar images being missed because of inaccurate and incomplete labeling of the older registered trademarks. Secondly, the VC system includes 29 categories, 145 divisions, and 1771 sections [3], yet it still cannot cover all possible shapes of trademarks. Another limitation is that the search results based on VC are not sorted by similarity, requiring examiners to check all found results, consuming a considerable amount of time and effort.

For these reasons, several leading countries have begun researching and experimenting with TIR using machine learning models or by combining them with VC search methods [4,5]. However, building a trademark image classification model poses several differences and challenges compared to regular image classification models. First, trademark images are often graphics-based and not directly captured photographs, making them prone to copying, assembling, and editing. These images usually consist of a few dominant colors but vary significantly in structural shapes. Another difficulty is that the trademark classification problem often lacks specific class labels, and training data labels typically involve pairs or groups of similar images. Additionally, evaluating similar images is also relative, and finding a comprehensive set of similar images with a given trademark in a vast trademark database is challenging. Moreover, trademark images often include both the protective graphic and textual components in the input image (see Fig. 1b). Many trademarks have protective graphic elements that occupy a very small area within the entire trademark image. Accurately identifying the protective graphic component is crucial to improving the accuracy of the classification model. However, in practice, image data in most countries usually store entire trademark images without specifying the exact region containing the protective graphic component [6], leading to challenges in constructing the training dataset.

(a) Trademark images consisting of graphics only

(b) Trademark images consisting of both text and graphic

Figure 1: Trademark image samples in the Vietnamese IP Office trademark dataset

One of the biggest challenges in TIR is the extremely limited amount of labeled data available for training Deep Convolutional Neural Network (DCNN) models, like Siamese Neural Networks (SNNs). Public datasets like METU [7] and USPTO contain around one million images but lack the necessary similar/dissimilar pairs or classification labels for training SNNs. The ability to retrieve images at a high conceptual level of similarity remains limited due to the lack of training data for such cases, a limitation that cannot be addressed through augmentation transformations. Our research aims to fill this gap by proposing a method to automatically generate a large labeled dataset using the national trademark dataset instead of public ones to train DCNN models. We employ Siamese and Triplet NNs, which are well-suited for learning similarity metrics from paired or triplet data. SNNs consist of two identical subnetworks that share weights and are trained to minimize the distance between similar pairs and maximize it for dissimilar pairs [8]. Triplet networks extend this concept by considering three inputs: anchor, positive, and negative, aiming to make the anchor closer to the positive than to the negative. These architectures have shown promise in various image retrieval tasks, and we adapt them to the specific requirements of TIR. Our approach introduces

the first key distinction from previous studies, which trained the model using a publicly available and pre-labeled dataset. Second, the enhanced architecture of the DCNN, coupled with an optimized loss function, significantly improves accuracy. As a result, our method effectively reduces both the time and effort required for evaluators, achieving a substantial efficiency gain.

Specifically, our work concentrates on four key innovative contributions, as follows. Firstly, we introduce a novel approach that enhances the effectiveness of TIR by incorporating additional structured information alongside trademark images. Secondly, we propose a method for constructing a suitable training and testing dataset through the automated generation of similar and dissimilar trademark data pairs, eliminating the need for manual labeling. Thirdly, we contribute a distinctive backbone model design for similarity comparison, integrated into both the Siamese and Triplet architectures. Another contribution is an adjustment to the Contrastive loss function (CLF) for SNN to better suit the specific characteristics of trademark data. This adjustment addresses challenges related to incorrect image pair labels, ensuring more accurate learning during training. As a result, the model achieves superior performance (NAR 0.0169 on the METU dataset), thereby minimizing subjective errors and manual processing time.

We structure the remaining sections as follows. Section 2 reviews the related research on machine learning and deep learning techniques for TIR, focusing on Siamese and Triplet models used for matching and extracting similar images in various fields, including trademark recognition. Section 3 describes the national dataset and outlines the procedures for building Siamese and Triplet models. Section 4 details the experimental setup and presents the results. The conclusions, limitations, and future research development are indicated in the final section.

2 Related Works

TIR is a class of Content-Based Image Retrieval (CBIR), specifically applied to trademark images. Over the past decades, various works have developed various schemes for CBIR [2,9]. Studies focused on using traditional techniques for extracting global features, such as color, edge, boundary, shape, spatial information, and texture [10–13]. These methods could effectively retrieve similar images invariant to transformations like translation, color variation, and scaling. However, they have not fully exploited local features containing crucial information for retrieving partially similar image instances. This perspective has also been partially highlighted by Li et al. [2].

To incorporate local features, subsequent research has extensively employed key-point-based methods, such as SIFT [14], SURF [15], LBP [16], and HOG [17]. These methods have enhanced the ability to retrieve similar images invariant to rotation and scaling transformations and to detect important and distinctive points in the images. As a result, they improved accuracy compared to earlier approaches. Nevertheless, these approaches exhibit constrained performance due to the restricted expressiveness and computational complexity associated with traditional feature extraction. Additionally, key-point-based methods still faced limitations in retrieving similar images at a high conceptual level of similarity, where similar trademarks might not have any common key points. Therefore, their practical use remained constrained.

With the explosive development of machine learning and especially various DCNN architectures [18–21], studies applying DCNNs for trademark similarity search have become the dominant direction. The results of these studies have shown a significant improvement in the effectiveness of DCNNs over traditional methods. Tursun et al. [18] utilized a pre-trained VGG16 model and an Extremal Region detector to locate and remove text components from images. Perez et al. [22] introduced a method for training two separate VGG19 models: one for learning visual similarities and the other for learning conceptual similarities. These models were trained on two datasets: one consisting of 151 classes, each containing 15 similar images collected from the web, and the other containing images and VC codes from a sub-dataset of USPTO [23]. This study

achieved an NAR of 0.066. This performance could be further improved by effectively leveraging local data and making adjustments in the training model. Tursun et al. [24] proposed using hard attention (ATRHA) and soft attention (SSA/CAMSA) methods to automate the removal of text components in images. These attention components were integrated into the model, with feature aggregations applied through MAC and R-MAC. This was also the best-performing model, achieving an NAR of 0.040, as synthesized by Li et al. [2] in the analysis of advanced TIR models on the METU dataset. The model's shortcomings could be addressed by utilizing advanced end-to-end deep learning architectures and refining pre-trained deep models through soft attention mechanisms. Tursun et al. [25] enhanced the R-MAC pipeline with three modifications: multi-resolution (MR), sum and max pooling (R-SMAC), and unsupervised attention (UAR). This study achieved remarkable results, with an NAR of 0.028 and a mAP@100 of 31.0%. Vesnin et al. [26] utilized a BEiT Vit model fine-tuned on the ImageNet-21k dataset, combined with PCAW/aQE/reranking techniques and local features, achieving the highest mAP@100 of 31.23%, but NAR was not reported. Bernabeu et al. [8] proposed a multi-label deep learning approach for trademark image retrieval (TIR), achieving the SOTA NAR of 0.018. The study did not describe the selection of the optimal weight ratio independently from the METU test set.

Apart from publicly available trademark datasets, some studies have developed recognition models based on mobile phone-captured datasets specifically for wine. Wine label image retrieval presents two key challenges. Firstly, the vast quantity of wine label images spans numerous brands, with varying sample sizes across different brands. Secondly, significant variations exist among wine label images of the same brand, whereas some labels from different brands exhibit only subtle distinctions. Lim et al. [27] developed an edge-histogram combined MLP model for wine label extraction, achieving an accuracy of 97.5% on a dataset of 517 samples. However, the system performed well only when the text on the wine labels was in English. Additionally, detecting candidate text regions using edge-based methods was often inaccurate in cases where font styles varied and character sizes differed significantly. The challenges of previous studies were solved by proposing a CNN-SURF Consecutive Filtering and Matching (CSCFM) framework [28,29]. The model achieved an average accuracy of 88.3% and a standard evaluation metric of 3.92 on the Oxford5k and UKB databases. The Transformer-based model, developed and tested on the Microsoft COCO (MS-COCO) and Flickr30K datasets, also demonstrated improved accuracy compared to previous studies [30]. Furthermore, a deep learning architecture integrating multimodal fusion was proposed to minimize the discrepancy between target and retrieved images within sensitive, domain-specific datasets [31]. Trappey et al. [32] proved that applying the advanced convolutional neural network model (VGG19) with a novel transfer learning approach enabled the test set to achieve a Recall@10 of 95%. The study emphasized that a significant limitation of the model could be the lack of diversity and quantity of images within the same category. Consequently, the sample selection process in the sampling strategy remains constrained. Future research should prioritize enhancing the organization of the training dataset and increasing data diversity to significantly improve the model's performance in retrieving visual semantic similarity.

In recent years, Siamese and Triplet architecture models have gained attention in many CBIR, tracking, identification, verification, and image object comparison studies [33,34]. SNN is an NN architecture containing two or more identical subnetworks with the same configuration, parameters, and weights. SNN-based algorithms currently dominate research in various fields [35] due to several notable advantages. Firstly, training SNNs only requires labeled similar/dissimilar pairs, which is often easier to implement than assigning classification labels to each image. Secondly, its unique learning mechanism sets SNN apart, allowing seamless integration with other conventional classifiers. This often leads to superior outcomes. Lastly, SNN focuses on considering features at deeper layers, strategically positioning similar features in proximity. Consequently, it demonstrates an aptitude for discerning aspects of semantic similarity within input data.

To our knowledge, some notable scientific studies have implemented SNN and Triplet networks to retrieve similar trademark images. Using the METU public dataset, Lan et al. [36] proposed a CNN model that combines Siamese and Triplet architectures, extracting handcrafted features from the convolutional feature maps. Trappey et al. [37] employed SNN to develop advanced models for assessing the similarity of trademark spelling, pronunciation, and images. Tursun et al. [38] introduced an approach using a Triplet NN to optimize a learned ensemble of Test-time Augmentation (TTA). The model was trained on 317 groups of similar trademarks, each containing at least two images. The study achieved impressive results across various test datasets, reaching a mAP@100 of 30.5% on the METU dataset. This model has yet to fully leverage local features that contain essential information for retrieving partially similar image instances.

Based on our review, we observe that most studies on TIR rely on publicly available datasets or relatively small self-constructed labeled datasets [22,24,26,36,38]. As mentioned in Section 1, the lack of a large-scale labeled dataset has limited the effectiveness of applying SNN models to TIR. By utilizing our proposed approach to create a large labeled dataset from the National IP Office database, our proposed model outperforms the SOTA NAR. This advancement demonstrates the great potential of applying national trademark data to enhance the model's ability to retrieve trademarks with higher accuracy and efficiency.

3 Methods

3.1 Trademark Database Description

This paper utilizes a dataset of Vietnamese trademarks maintained and managed by the IP Office of Vietnam. The dataset comprises over 750,000 trademark applications in Vietnam and more than 150,000 international trademark applications designated for protection in Vietnam as of the end of 2023. The registered trademarks accepted for protection are published by the IP Office of Vietnam [39]. They utilize the Oracle database management system, developed based on the WIPO IPAS (World Intellectual Property Organization Industrial Property Automation System), to store trademark applications and registrations. This volume of data is sufficient to train large data models and artificial intelligence systems. The database includes comprehensive fields that describe trademarks, the application process, and examination results, which can be leveraged for various problem-solving tasks.

Since the evaluation process involves examiners assessing images using the VC system, there are no pre-existing lists of trademarks labeled as similar in the Vietnamese trademark database, akin to publicly available labeled trademark datasets, such as the METU dataset [7]. However, the Vietnamese dataset offers additional information in several data fields that are not present in global public datasets, including:

- Data on trademark application rejections due to similarity with already protected trademarks.
- Information on the citation dataset for trademark applications and registrations.
- Information on the duration and history of trademark protection, including expiration dates.
- Information on trademarks that have been protected by companies over time.

3.2 Data Processing and Augmentation

The national trademark database comprises both textual and image trademark applications, and the citation list includes both other textual and image trademarks. A "citation list" is a compilation of trademark applications or registrations that share similarities with the trademark application under examination and were filed before it. This list is generated by the examiner during the evaluation process. To ensure effective preprocessing, we implement the following procedures:

- Selection of image-based trademarks: We retain only those trademarks that have visual components of interest to trademark protection. This is a crucial step because it concerns itself with visual trademarks

only that consist of trademark images comprising graphics only and trademark images comprising text and graphics.

- VC filtering: We further filter the dataset by selecting only those image-based trademarks that have an associated VC.

- Curating relevant citation: In the citation list, we retain only those trademarks with images that share at least one VC in common with the trademarks under examination.

The data processing workflow consists of critical steps, including image enhancement, image resizing to 224 × 224 pixels, and background removal. To improve the model's robustness and prevent overfitting, we apply data augmentation procedures in the training procedure, including random cropping, scaling, flipping, rotation, translation, color jittering, brightness/contrast, adding noise, erasing, and blurring.

3.3 Proposed Training Dataset Construction

Leveraging the dataset including additional data fields of VC details, application owner information, and citation list, we develop algorithms to automatically match trademark images. To automatically generate similar image pairs, we propose using the following four methods (Algorithm 1). By incorporating a weighted balancing mechanism, we ensure that the training dataset maintains a harmonious distribution of similar and dissimilar image pairs across different methods. This prevents any single method from disproportionately influencing the dataset and enhances the overall quality and robustness of the model. Creating suitable pairs of similar/dissimilar images for input to Siamese architectures is necessary. For the Triplet architecture, it is required to provide triplets consisting of anchor, positive, and negative samples as input. To generate triplets for the Triplet model, we first select a random pair of data that may be similar or dissimilar, and then find the remaining trademark for the triplet using feasible methods (as some trademarks may not have other trademarks with the same owner, the citation list may lack suitable trademarks, or some VC codes may not have corresponding trademarks). Due to this limitation in creating image triplets, the training dataset is constructed with pairs, which are more suitable than triplets, making it more suitable to construct the model using a Siamese architecture rather than a Triplet architecture. The process of creating a dataset of brand image pairs for training the system is displayed in Fig. 2.

Algorithm 1: Automatic generation of similar trademark image pairs

Input: Trademark database D with images, citation lists, VC, and ownership information

Output: Selected similar pairs based on the chosen method

1. For each trademark $t \in D$:

- Add (t, t, "self-pairing") to S as a similar pair
- **For each citation** $c \in C_t$ (C_t is citation list of t):
  - If c has an image and shares at least one VC with t:
    - Add (t, c, "citation") to S as a similar pair
- **For each trademark** $o \in O_t$ (O_t is same owner list of t):
  - If o has an image and shares at least one VC with t:
    - Add (t, o, "owner") to S as a similar pair

2. Perform a weighted selection on S to obtain the final set of similar pairs.

3. Apply real-time augmentation to images in selected similar pairs (e.g., rotation, cropping, color adjustment, brightness modification).

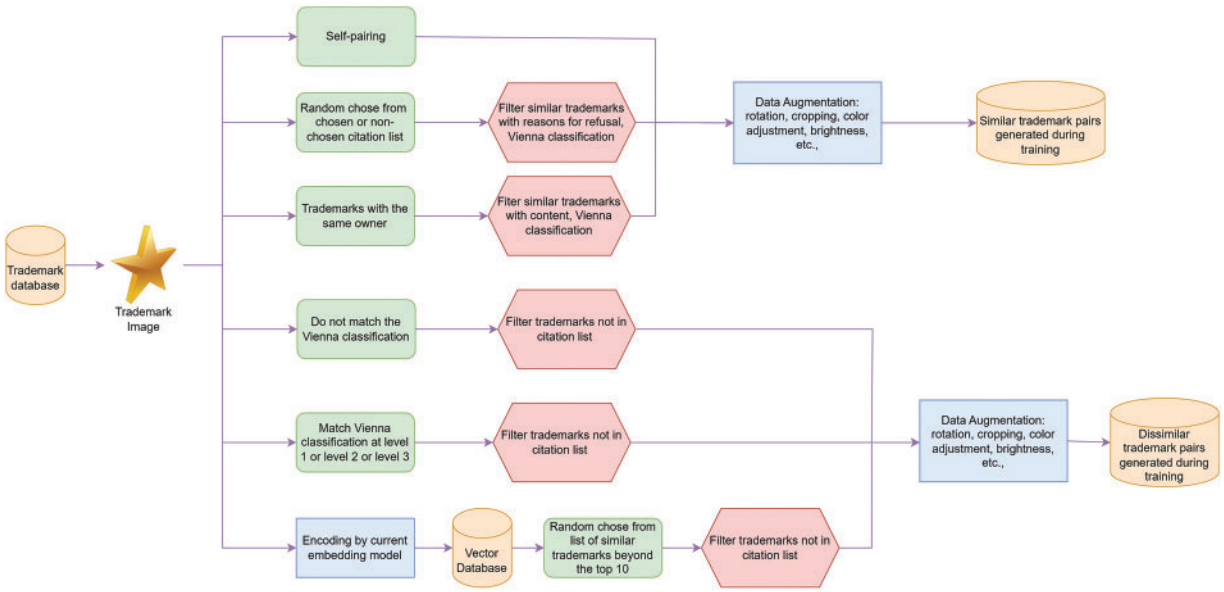

Figure 2: Schematic of creating trademark image pairs dataset for training

3.4 Model Training Architecture

3.4.1 Siamese Architecture

An SNN is a specialized NN architecture designed to work with pairs of input data [40]. The primary purpose of SNNs is to measure the similarity between two input data points by comparing the feature vectors that are extracted by the subnetworks (see Fig. 3). The CLF is commonly used in SNNs to guide the learning process. This loss function quantifies how well the network distinguishes between similar and dissimilar pairs.

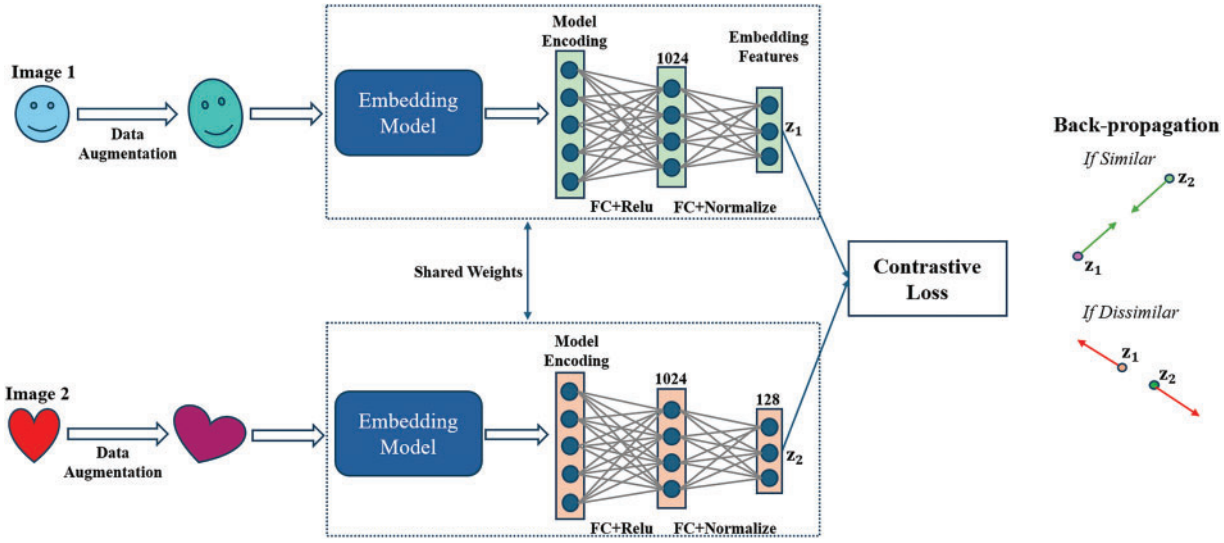

Figure 3: Training process using a Siamese architecture model

Mathematical formulation: In [41], the general Contrastive loss can be defined as:

$$\mathrm{L}(W) = \sum_{i=1}^{p} L\left(W, \left(Y, \vec{X}_1, \vec{X}_2\right)^i\right) \tag{1}$$

$$L\left(W, \left(Y, \vec{X}_1, \vec{X}_2\right)^i\right) = (1-Y)\, L_S\left(D_W^i\right) + Y L_D\left(D_W^i\right) \tag{2}$$

$$D_W\left(\vec{X}_1, \vec{X}_2\right) = \left\| G_W\left(\vec{X}_1\right) - G_W\left(\vec{X}_2\right) \right\|_2 \tag{3}$$

where: L is the loss function for a sample; $\vec{X}_1, \vec{X}_2$ are input vectors; G_W represents the characteristic function of the NN, which transforms input vectors into embedding features through the network; W are the learnable parameters of the parameterized function G_W; Y is a binary label established for this pair with:

$$Y = \begin{cases} 0 \text{ if } \vec{X}_1, \vec{X}_2 \text{ are deemed similar} \\ 1 \text{ if } \vec{X}_1, \vec{X}_2 \text{ are deemed dissimilar} \end{cases} \tag{4}$$

$\left(Y, \vec{X}_1, \vec{X}_2\right)^i$ is the i-th sample of training pairs; L_S is the partial loss function for a similar sample; L_D is the partial loss function for a dissimilar sample; P is training pairs; D_W is the distance function as the Euclidean distance between output embeddings $G_W\left(\vec{X}_1\right)$, $G_W\left(\vec{X}_2\right)$.

The exact CLF:

$$L_S\left(D_W\right) = \frac{1}{2}\left(D_W\right)^2 \tag{5}$$

$$L_D\left(D_W\right) = \frac{1}{2}\left\{\max\left(0, m - D_W\right)\right\}^2 \tag{6}$$

$$\Rightarrow L_W\left(Y, \vec{X}_1, \vec{X}_2\right)^i = (1-Y)\frac{1}{2}\left(D_W^i\right)^2 + (Y)\frac{1}{2}\left\{\max\left(0, m - D_W^i\right)\right\}^2 \tag{7}$$

where $m > 0$ is a margin value that defines the minimum distance between dissimilar pairs [41].

3.4.2 Triplet Architecture

A Triplet NN is a type of NN architecture specifically designed to learn how to differentiate between similar and dissimilar data points by using a set of three inputs: an anchor, a positive sample, and a negative sample [42]. The Triplet NN structure comprises three primary branches, including Anchor (A), Positive (P), and Negative (N) (see Fig. 4). The goal of the Triplet loss is to minimize the distance between the anchor and positive pair while maximizing the distance between the anchor and negative pair. This process refines the model's ability to accurately distinguish between similar and dissimilar samples.

Mathematical Formulation: The Triplet loss can be defined as:

$$\mathrm{L}_{\text{triplet}}(A, P, N) = \max\left(d(A, P) - d(A, N) + \alpha, 0\right) \tag{8}$$

where A, P, N are three data points: anchor, positive, and negative; $d(A, P)$ is the distance between the anchor and the positive in the representation space; $d(A, N)$ is the distance between the anchor and the

negative in the representation space; α is a margin value used to ensure that the distance between anchor and positive is smaller than the distance between anchor and negative before the loss is computed.

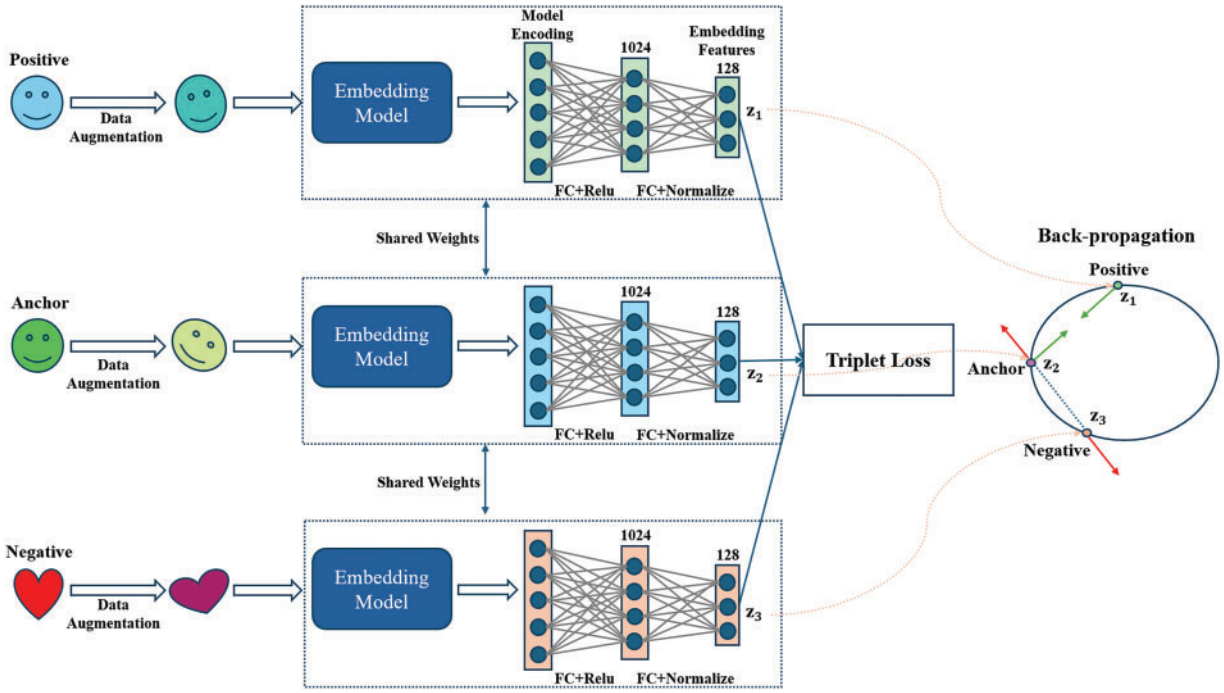

Figure 4: Training process using a Triplet architecture model

3.5 Backbone Model Design

The backbone model integrates the embedding model and additional layers, enhancing overall performance. The two main components are:

- Embedding Model: The embedding model is essential in extracting features from input images. It captures various aspects such as color, shape, and texture. In this study, we evaluate multiple embedding models, each differing in architecture and model size. This allows for a comprehensive comparison of embedding effectiveness. The selected architectures, ResNet, VGG-19, EfficientNet, and Vision Transformer (ViT), have been shown to perform well on various image datasets. The approach identifies a wide range of features and allows one to study how various model architectures influence embedding performance.

- Fully Connected (FC) Layers: The output features from the embedding model undergo refinement through two FC layers:

- The first FC layer (1024 features, ReLU activation) performs a non-linear transformation of the embedding features. The selection of 1024 features in the first FC layer allows for a higher-dimensional space where complex relationships among the features can be captured. This higher dimensionality helps to preserve the richness of the extracted features from the embedding model, enabling more intricate patterns and characteristics to be learned.
- The second FC layer (128 features, normalized activation) stabilizes and scales feature values, significantly improving the training process. This is particularly beneficial in models like Siamese and Triplet, where consistent and balanced features are crucial for similarity comparison tasks. Normalization

ensures well-distributed feature values, enhancing the model's ability to learn accurate similarities and differences between input pairs.

The refined features from each image pair are used to train the embedding model and FC layers using Contrastive Loss or Triplet Loss. This design incorporates established embedding models and critical layers, optimizing feature learning for similarity comparison tasks.

3.6 Adapted Contrastive Loss Function (ACLF) for Trademark Data

As seen in Eqs. (5)–(7), we observe that from the similar-loss function L_S, the backpropagation process ensures that the two output embeddings $G_W\left(\vec{X}_1\right)$ and $G_W\left(\vec{X}_2\right)$ converge as closely as possible when $\vec{X}_1$ and $\vec{X}_2$ are deemed similar. Conversely, the dissimilar-loss function L_D will cause $\vec{X}_1$ and $\vec{X}_2$ to diverge further apart when $\vec{X}_1$ and $\vec{X}_2$ are dissimilar and the distance D_W is less than m. When D_W exceeds m, L_D equals zero, and the model will stop pushing $\vec{X}_1$ and $\vec{X}_2$ apart.

We introduce the Adapted Contrastive Loss Function (ACLF) for SNN. The idea behind our modification stems from two main reasons. First, we observe that labeled datasets automatically created from citation lists contain significant noise due to incorrect labeling. This is unavoidable because many trademarks in citation lists overlap with the trademark under examination in the text, while the images are completely different. Moreover, since citation lists are built over decades with hundreds of different examiners, errors may inevitably occur during the examination process. If the original CLF formula is used, the model's accuracy will be significantly affected by these incorrect trademark pairs.

The second reason arises from the specific characteristics of TIR compared to other image comparison models. In TIR, specifically in finding citations for trademarks, we observe that a pair of data points is considered a citation if they share an essential part of the image that differentiates them. It is possible for trademark A to be similar to trademark B in one aspect, and B to be similar to trademark C in another aspect, while A and C are completely different with no common features. In the original formulation, L_S makes $G_W\left(\vec{X}_A\right)$ and $G_W\left(\vec{X}_B\right)$ converge to a distance of 0, and $G_W\left(\vec{X}_B\right)$ and $G_W\left(\vec{X}_C\right)$ also converge to 0. Consequently, $G_W\left(\vec{X}_A\right)$ and $G_W\left(\vec{X}_C\right)$ will also converge to 0 even though A and C are completely different.

To address this limitation, we propose ACLF with the following formulation:

$$L_S\left(D_W\right) = \frac{1}{2}\left\{\max\left(0, D_W - k\right)\right\}^2 \tag{9}$$

$$\Rightarrow L\left(W, \left(Y, \vec{X}_1, \vec{X}_2\right)^i\right) = (1-Y)\frac{1}{2}\left\{\max\left(0, D_W^i - k\right)\right\}^2 + (Y)\frac{1}{2}\left\{\max\left(0, m - D_W^i\right)\right\}^2 \tag{10}$$

with:

$$D_W\left(\vec{X}_1, \vec{X}_2\right) = \left\|G_W\left(\vec{X}_1\right) - G_W\left(\vec{X}_2\right)\right\|_2 \tag{11}$$

$$Y = \begin{cases} 0 \text{ if } \vec{X}_1, \vec{X}_2 \text{ are deemed similar} \\ 1 \text{ if } \vec{X}_1, \vec{X}_2 \text{ are deemed dissimilar} \end{cases} \tag{12}$$

The gradient of L_S and L_D:

$$\frac{\partial L_S}{\partial W} = \begin{cases} (D_W - k)\frac{\partial D_W}{\partial W} & \text{if } D_W > k \\ 0 & \text{if } D_W \leq k \end{cases} \tag{13}$$

$$\frac{\partial L_D}{\partial W} = \begin{cases} -(m - D_W)\frac{\partial D_W}{\partial W} & \text{if } D_W < m \\ 0 & \text{if } D_W \geq m \end{cases} \tag{14}$$

Fig. 5 visualizes the components of the Contrastive similar-loss function L_S and dissimilar-loss function L_D in the original formulation, while Fig. 6 visualizes the component of the adapted Contrastive similar-loss function in our modified formulation within the 2D Euclidean embedding space. From Eq. (10) and as illustrated in Fig. 6, it is evident that this modification to the L_S formulation will cause the vectors $G_W\left(\vec{X}_1\right)$ and $G_W\left(\vec{X}_2\right)$ to converge towards a distance of k instead of zero. If the distance between $G_W\left(\vec{X}_1\right)$ and $G_W\left(\vec{X}_2\right)$ is less than k, the model will no longer attempt to bring them closer together.

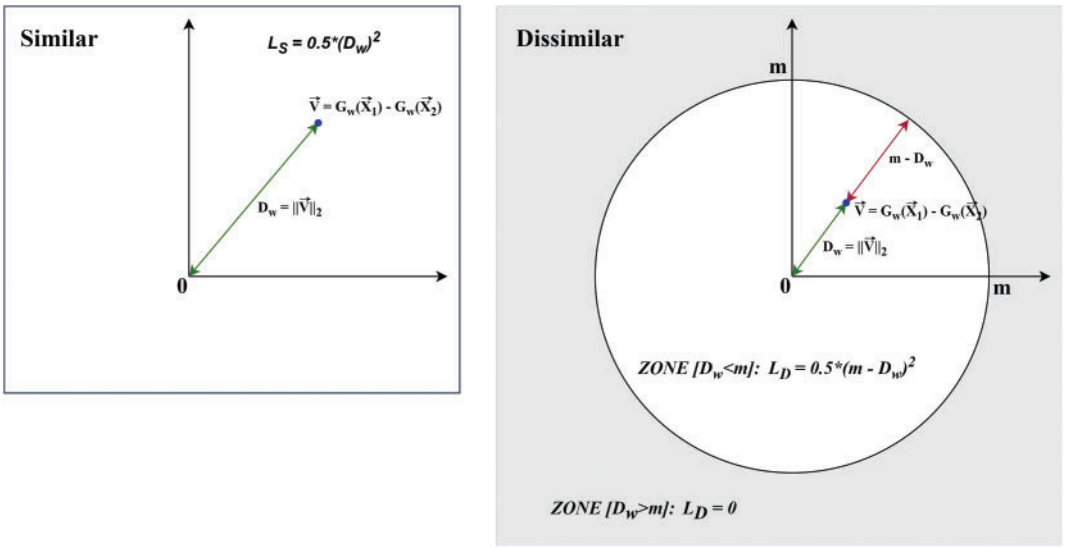

Figure 5: The Contrastive similar-loss function (L_S) and dissimilar-loss function (L_D) according to the original formulation in the 2D Euclidean embedding space

Comparing both the similar loss function (Eqs. (5) and (9)), we modify the D_W component by introducing $\max(0, (D_W - k))$. This adjustment shifts the optimization target for D_W from absolute zero to a predefined threshold k. This modification is conceptually similar to the L_D term in the original Contrastive Loss but, instead of restricting D_W from exceeding a margin m, we introduce a lower bound k, ensuring that D_W does not become arbitrarily small. This modification prevents the model from overly compressing similar pairs, allowing for a more nuanced representation of similarity. Furthermore, enforcing a minimum distance between similar pairs makes the model less sensitive to noisy labels.

To illustrate the impact of this adjustment, we present Fig. 7, which visualizes the embedding space before and after training with the ACLF. When using the original CLF, PCA Component 1 and 2 values are primarily concentrated within the range $(-0.5, 0.5)$. This occurs because L_S attempts to minimize the distance between images within the same group to near zero, reducing the available degrees of freedom. However, due

to some labels belonging to multiple citation lists and the presence of mislabeled noisy samples, L_S not only pulls together labels from the same citation list but also causes entire citation list groups to converge. While the L_D term is designed to push different groups apart, it becomes inactive (i.e., zero) when $D_W > m = 1$, leading to all embeddings collapsing into a circular region with a diameter of 1.

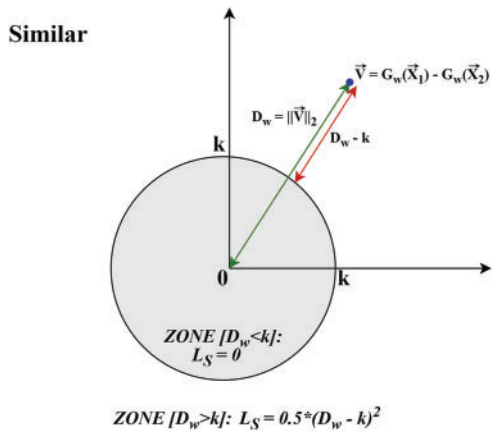

Figure 6: The adapted Contrastive similar-loss function in the 2D Euclidean embedding space

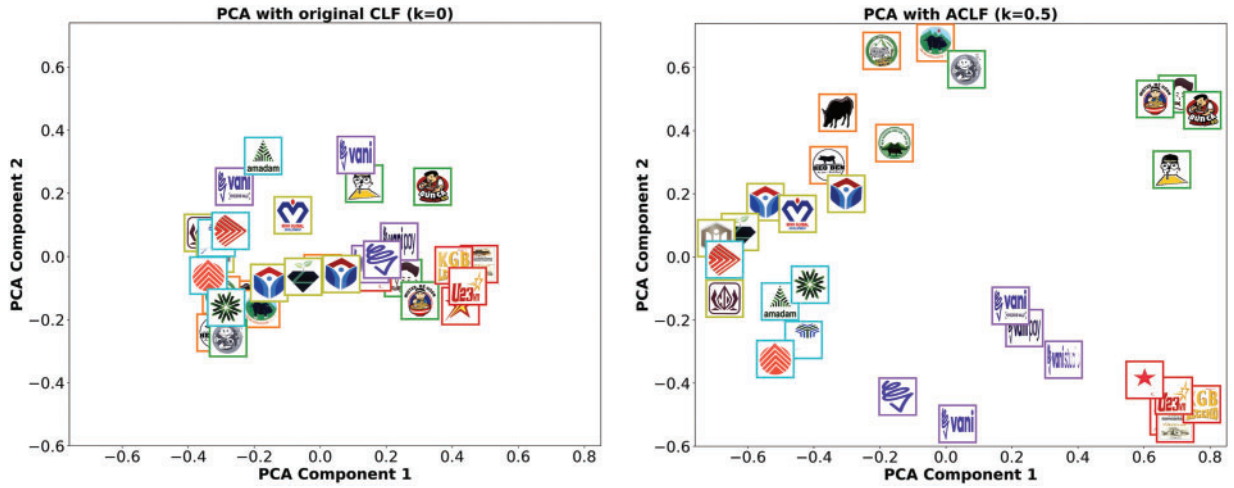

Figure 7: Comparison of original CLF and ACLF in PCA visualization (same color denotes the same citation list)

In contrast, applying the ACLF results in a significantly expanded PCA distribution. This modified loss function reduces the attraction force between different groups and increases the flexibility of the parameter space, preventing D_W from collapsing to zero. While intra-group distances slightly increase, the separation between different citation list groups becomes much more distinct.

3.7 Similarity Measurement

To evaluate our models, we split the Vietnamese trademark dataset into a training set (80%) and a validation set (20%). This split ensures sufficient data for training while maintaining a large enough validation set to reliably assess model performance. To ensure fairness of the evaluation, after splitting the dataset, we

perform additional data processing: in the citation list of samples in the training dataset, we remove citations that belong to the validation dataset.

From the training dataset, we generate pairs of data to serve as input for the Siamese model and triplets for the Triplet model, using the methodologies outlined in Section 3.3. These models are then trained using either the CLF or the Triplet loss function.

To compare and evaluate the effectiveness of the Siamese and Triplet models, we only create similar pairs on the validation dataset, then compute the embeddings of the images and predict the labels Y based on a cut-off threshold. Additionally, in this evaluation, similar pairs are exclusively generated using the "chosen citation pairs" and "non-chosen citation pairs" methods, while dissimilar pairs are created using the "random pairs with VC level 3" method. We use citation list pairs for similar images because citations represent the key trademarks examiners seek. Dissimilar pairs are selected using the "random pairs with VC level 3" method, as these pairs are harder to distinguish due to shared VC similarities, while other methods may ease recognition, inflating performance metrics that do not align with the practical goals of examiners.

Additionally, since the models are trained with different loss functions, the cut-off thresholds for the models vary significantly. Therefore, instead of adhering to the conventional 0.5, we select additional metrics to evaluate the model's performance. From an overview of other TIR works, we observe that the two most commonly used metrics for evaluating results are NAR (Normalized Average Rank) and mAP@k (mean Average Precision of the top k results). The mAP@k metric is calculated using the following formula:

$$mAP@k = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{E}\sum_{j}^{k}\frac{r_j}{j} \tag{15}$$

where Q is the number of queries; E_i is the number of expected results of the query i; r_j is similar-rank of the image when the image with rank j is the expected image, and otherwise is zero.

The formula for NAR is as follows:

$$NAR = \frac{1}{N \times N_{rel}}\left(\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}\left(N_{rel}+1\right)}{2}\right) \tag{16}$$

where N_{rel} is the number of relevant images, R_i is the rank of the i-th relevant injected image, and N is the total size of the image dataset.

We observe that using NAR and mAP@k to evaluate the National IP Office trademark dataset has significant drawbacks. First, with mAP@k, given by Eq. (15), this metric is heavily influenced by the order of a few correctly retrieved images at the top of the results. However, for assisting examiners, a TIR system requires a very high recall to avoid missing too many cases. This is the reason we exclude mAP@k as a performance evaluation metric.

Next, with Eq. (16), NAR reflects the ability to evaluate the average ranking of retrieved similar images. However, this also makes NAR heavily dependent on the order of the least similar images. When applied to the labeled dataset, which contains a significant number of noisy labels, this dependence leads to inaccurate evaluations. Therefore, while NAR is relatively effective for assessing the carefully curated query dataset of METU, it cannot be reliably applied to our automatically generated dataset.

After considering various metrics, we evaluate model performance based on TPR at fixed and very high TNR thresholds of 0.9 and 0.95. This decision stems from the practical reality that the number of true positive (TP) labels—similar trademark images—is extremely small compared to the number of true negative (TN) labels—dissimilar trademark images. Setting a high TNR threshold helps to significantly reduce the number of incorrect labels that examiners need to review. At a TNR threshold of 0.95, examiners still need to review

5% of the dissimilar images predicted as similar, amounting to tens of thousands of labels. While this is a substantial number, requiring additional filtering measures for practical applications, it represents a step toward feasibility.

Another advantage of evaluating TPR at high TNR thresholds is that it remains relatively stable even when the number of negative labels increases, as the number of positive labels typically stays constant. This situation is common since newly protected trademarks are required not to overlap with existing ones. In contrast, metrics like mAP@100 are likely to decline significantly as the number of negative labels grows rapidly while the number of positive labels remains unchanged. The choice of a high TNR threshold (0.9, 0.95) is justified by the significantly larger number of applications with differing labels compared to those with matching labels. A high TNR helps efficiently filter out many irrelevant applications, which is crucial for practical applications.

4 Experimental Results and Evaluation

4.1 Experimental Environment

To experiment with the proposed methods and algorithms, we use the following testing environments: CPU 24 cores, 128 GB RAM, 2 GPU A100; Ubuntu server 22.04; Search Engine and Vector Database: Elastic Search 8.8.1; Programming Language: Python 3.9.

4.2 Data Processing Results

We find that only 332,972 samples, representing 35.3% of the trademarks (see Table 1), include both images and VC codes, making them suitable for inclusion in the training and testing dataset. We then split the data into training and testing sets with an 8:2 ratio. After eliminating entries that lack counterparts or whose counterpart lists include only trademarks without images, there are 198,270 trademarks in the training dataset (21% of the total number of trademarks) and 51,474 trademarks in the validation dataset (5.5% of the total number of trademarks).

Table 1: Statistics on image-based trademarks within the trademark dataset in the Vietnamese IP Office

Description	Quantity	Proportion (%)
Trademarks without images	510,829	54.2%
Trademarks with images but without Vienna codes	99,504	10.5%
Trademarks with images and Vienna codes	332,972	35.3%
Total number of trademarks	**943,305**	**100%**

Subsequently, we perform background removal utilizing the Rembg tool based on U^2-Net [43], enhance the images, and resize them to dimensions of 224 × 224 pixels. Fig. 8 illustrates some pairs of original image data and the processed image data. Through the processed trademark images, we observe that the background and text components are effectively removed, retaining only the dissimilar areas of the images that need to be considered for protection.

Figure 8: Sample pairs of original trademark images and images after background removal and enhancement

4.3 Trademark Image Pairs

After obtaining a suitable training and validation dataset, we develop functions to create pairs of similar image pairs and dissimilar image pairs according to the methods proposed in Section 3.3. Among the methods for constructing trademark image pairs, the two most important and meaningful for examiners are the methods for selecting similar image pairs: chosen citation pairs and non-chosen citation pairs. These represent the citation trademarks that examiners most want to identify. Our training pair dataset includes 13,144 chosen citation pairs and 490,355 non-chosen citation pairs, generated from 198,270 distinct images. The validation pair dataset includes 4090 chosen citation pairs and 154,480 non-chosen citation pairs.

From the images of trademarks in the chosen citation list and the non-chosen citation list presented in Fig. 9, we observe that the pairs of images generated using this method are very suitable for training the SNN and Triplet NN models. This is because the contrastive image pairs are both diverse and not entirely identical to the original images, yet the main protective components share enough similarity to serve as valid comparisons from the examiner's perspective. Although the citation list may still contain some images that differ significantly from the trademarks under examination due to various reasons, this proportion is minimal. We believe the citation list is the most important component for enabling the model to learn to identify pairs of images that are considered similar in the field of trademark protection.

4.4 Pre-Trained and Non-Pre-Trained Model Evaluation

With nearly 200,000 trademark images in the training set, we assess this sample size as average-sufficient for training models but relatively low compared to other image datasets, the most notable being ImageNet-1k, which contains 1,281,167 training images across 1000 classes. While the images in ImageNet-1k are photographs, those in the trademark dataset are primarily graphic images designed by humans. Nevertheless, models can inherit many features from pre-trained models.

Therefore, we conduct experiments and evaluations of embedding models trained from scratch as well as models fine-tuned from pre-trained models based on the ImageNet-1k dataset. The evaluation results are presented in Fig. 10 and Table 2; for the models trained from scratch, we add the symbol/n at the end of the model's name.

From the experimental results, we find that among the models trained from scratch, the VGG19bn Siamese model performs better than both the ResNet50 Siamese and ResNet50 Triplet. The VGG19bn_Siamese/n model achieves a TPR of 0.447 at a TNR of 0.9, improving by 13.4% on ResNet50_Triplet/n and 15.5% on ResNet50_Siamese/n. Even those models trained from scratch fare much worse compared to the models that were fine-tuned using the pre-trained models of ImageNet-1k. For a TNR of 0.9, the pre-trained models exhibit an increase in TPR compared to scratch-trained models: 39.2% for ResNet50_Triplet, 43.3% for ResNet50_Siamese, and 26.2% for VGG19bn_Siamese. Based on these

evaluation results, we only fine-tune the models that have been pre-trained on the ImageNet-1k dataset in the subsequent experimental sections.

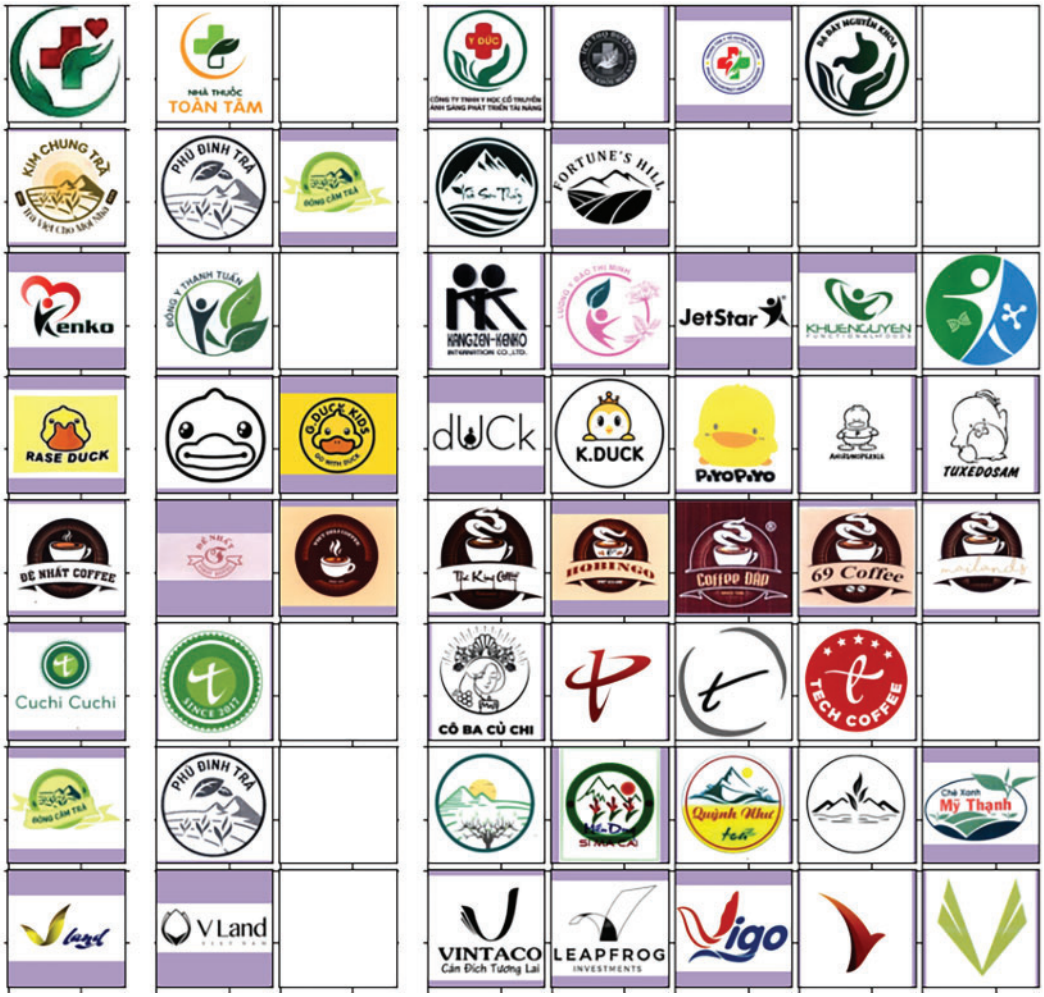

Figure 9: Sample pairs of original trademarks, top 2 trademarks from the chosen citation list, and top 5 trademarks from the non-chosen citation list

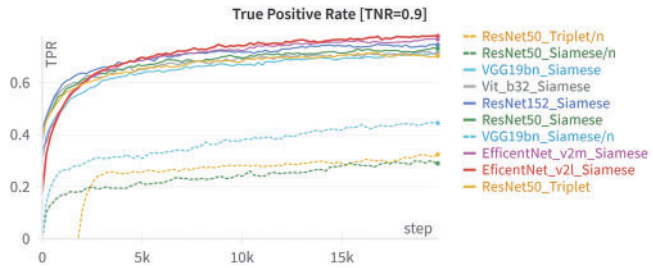

Figure 10: True positive rate comparison: training from scratch and fine-tuning from pre-trained ImageNet-1K

Table 2: Experimental results of the models with the validation dataset

Model	Pretrained	k	TNR = 0.9		TNR = 0.95		ROC_AUC
			TPR	Accuracy	TPR	Accuracy	
ResNet50_Triplet/n	No	–	0.313	0.607	0.225	0.563	0.68
ResNet50_Siamese/n	No	0.5	0.375	0.638	0.288	0.594	0.726
VGG19bn_Siamese/n	No	0.5	0.447	0.674	0.367	0.634	0.756
ResNet50_Triplet	Yes	–	0.705	0.803	0.618	0.759	0.883
ResNet152_Triplet	Yes	–	0.725	0.813	0.638	0.769	0.888
EfficentNet_v2m_Triplet	Yes	–	0.717	0.809	0.631	0.766	0.887
Vit_b32_Triplet	Yes	–	0.7	0.8	0.62	0.76	0.878
ResNet50_Siamese	Yes	0.5	0.725	0.813	0.646	0.773	0.895
ResNet152_Siamese	Yes	0.5	0.745	0.823	0.661	0.781	0.902
VGG19bn_Siamese	Yes	0.5	0.709	0.805	0.626	0.763	0.887
EfficentNet_v2m_Siamese	Yes	0.5	0.763	0.832	0.682	0.791	0.909
EficentNet_v2l_Siamese	Yes	0.5	**0.777**	**0.839**	**0.708**	**0.804**	**0.914**
Vit_b32_Siamese	Yes	0.5	0.714	0.807	0.624	0.762	0.89

4.5 ACLF Performance

As mentioned in Section 3.6, we propose the use of the ACLF according to Eq. (10) to better align with the task of finding similar trademarks. To test and evaluate the effectiveness of this new formulation, we experiment with the ResNet50_Siamese model using different values of the coefficient k from the set [0.0, 0.1, 0.3, 0.5, 0.7]. The coefficient m is set to the default value of 1.0, as specified in the original algorithm. Given that we perform normalization in the final neural layer, the Euclidean distance between the two output embedding vectors lies within the range [0, 2], making the choice of m = 1.0 appropriate.

The results of the experiments with the ResNet50_Siamese model using various k coefficients are illustrated in Fig. 11 and Table 3. From the experimental results, we observe that for $k > 0$, the outcomes are significantly better compared to $k = 0$ (where the ACLF is equivalent to the original unmodified Contrastive loss function). Table 3 indicates that the optimal coefficient k is 0.5, which resulted in an increase in TPR of 41.7% compared to $k = 0$. The experimental results also demonstrate that models trained with $k = 0.3$ and $k = 0.7$ perform very well, with only minor differences in TPR compared to the model trained with $k = 0.5$.

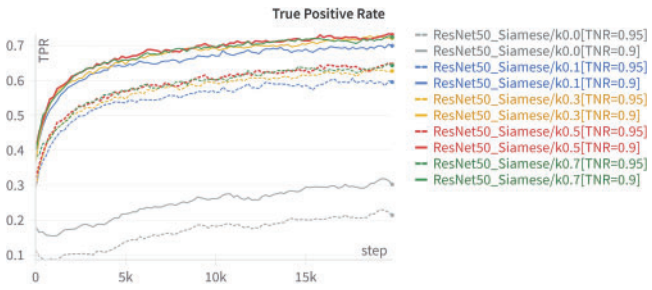

Figure 11: TPR for Siamese models with varying k-values using ACLF

Table 3: Experimental results of the ACLF with different *k* parameters

Model	Pretrained	k	TNR = 0.9		TNR = 0.95		ROC_AUC
			TPR	Accuracy	TPR	Accuracy	
ResNet50_Siamese/k0.0	Yes	0	0.308	0.604	0.215	0.558	0.661
ResNet50_Siamese/k0.1	Yes	0.1	0.703	0.802	0.596	0.748	0.878
ResNet50_Siamese/k0.3	Yes	0.3	0.722	0.811	0.627	0.764	0.892
ResNet50_Siamese/k0.5	Yes	0.5	**0.725**	**0.813**	**0.646**	**0.773**	**0.895**
ResNet50_Siamese/k0.7	Yes	0.7	0.722	0.811	0.643	0.772	0.887

4.6 Siamese Model vs. TRIPLET Model

We test the Siamese and Triplet models using the embedding models. The results of the experiments for the Siamese and Triplet models are presented in Table 2. From the experimental findings, we observe that for models fine-tuned from the pre-trained ImageNet-1k dataset, utilizing the Siamese architecture combined with the ACLF yields superior results compared to employing the Triplet architecture and Triplet loss function across all embedding models we examine.

In our view, the superior performance of the Siamese model compared to the Triplet model is driven by two factors:

- The ACLF is specifically designed for SNN architectures, whereas Triplet NN architectures do not require such adjustments. This modification significantly enhances the effectiveness of SNN models.

- The training dataset is more compatible with the Siamese architecture than with the Triplet architecture. This has been mentioned in Section 3.3.

As Siamese models paired with the ACLF yield better performance than Triplet models with identical backbone architectures, subsequent experiments focus on training Siamese models using diverse backbones to assess the effectiveness of each.

4.7 Compare Embedding Models

** The dataset of 521 images: 9 images from the citation list (marked with blue scores) and 512 other images with the same Vienna code 03.07.03, described as "Cocks, hens, chickens" (marked with red scores). The results are sorted by descending similarity scores.*

We employ the same Siamese architecture to train models utilizing various embedding models, each pre-trained from the ImageNet-1k dataset, including ResNet-50, ResNet- 152, EfficientNet-v2_m, EfficientNet-v2_l, VGG19_bn, and ViT_b_32, to assess their effectiveness. From the experimental results presented in Table 2, the models based on the EfficientNet architecture yield the best performance, followed by the ResNet models, with ViT_b_32 and VGG19_bn trailing behind. Notably, while VGG19_bn demonstrates superior performance compared to ResNet50 in models trained from scratch, its performance is lower when utilizing pre-trained weights. The EfficientNet_v2l_Siamese model achieves the highest performance at both cut-off thresholds of TNR = 0.9 and TNR = 0.95, with corresponding TPRs of 77.7% and 70.8% and accuracies of 83.9% and 80.4%, respectively. The model also exhibits the highest ROC_AUC of 91.4%.

Fig. 12 illustrates the results of a similar image search for a query image within the test dataset (the query image is the first image in Fig. 12a). The search results are ordered by descending similarity score (which ranges from [−1, 1]). In this example, we retrieve 6 out of 9 images from the citation list within the top 30 images and 9 out of 9 citations within the top 60 images. The three citation images found between the top

30–60 (two of which can be seen in Fig. 12b) show significant differences from the query image, whereas the 6 citation images within the top 30 are quite similar.

(a) Top 30 Most Similar Trademarks (First Image is the Query Image)

(b) Images with Similarity Scores Ranked 42 to Ranked 71

(c) Images with Similarity Scores Around Ranked 200

Figure 12: Result of searching for similar images (using model EfficientNet_v2l_Siamese)

Additionally, as shown in Fig. 12a, many images within the top 20 are visually similar to the query image, even though they are not part of the citation list. This demonstrates the model's effectiveness in searching and ranking results by similarity score, which can be leveraged to reduce the effort and time and improve the accuracy of examiners assessing new applications.

The similarity score of 0 is a particularly significant threshold because we set the parameter $m = 1$ in the CLF formula. During training, if the distance between the embeddings of two images exceeds 1 (equivalent to a similarity score of 0 in our calculation), the model stops pushing the two embeddings further apart, as the images are already different enough to confirm that the pair is dissimilar. As a result, when the similarity score drops below 0, the retrieved images are significantly different from the query image (similarity score <

0 begins from the 59th image out of 521). Thus, by using a similarity score threshold of 0 as a cutoff point, the number of images an examiner needs to compare in this query image example could be reduced by nearly 9 times.

4.8 Comparison with Other Studies

To compare our results with other studies, we test the EfficientNet_v2l_Siamese model on the METU_v2 trademark dataset. METU_v2 comprises two main subsets: a test dataset containing over 930,000 unlabeled trademark images and a query dataset consisting of 417 trademark images categorized into 35 similar groups. Following the standard testing protocol adopted in other studies [7], we inject the 417 query images into the unlabeled test dataset. Next, we sequentially query each image from the query dataset and compute the Euclidean distance between the query image and the images in the combined dataset. Then we rank the images within the same group of the query image and compute the Average Rank (AR), the NAR score.

Table 4 shows a comparison of our model's result with those reported in other studies when tested on the METU dataset. Our model achieves a NAR score of 0.0169, outperforming the current SOTA NAR. The AR of 15,498.4 shows that while our model outperforms the SOTA NAR and AR, additional filtering methods, such as those based on VC or time, are still necessary for effective real-world application. This is because, in practice, examiners usually prefer to review only a few hundred results or fewer. This highlights the need for future research to explore more effective solutions. Compared to the SOTA NAR achieved by Bernabeu et al. [8], both approaches incorporate additional metadata beyond images. However, they employed a hybrid model combining color (30%) and shape (70%) optimized for the METU dataset, which risks overfitting to a specific dataset. In contrast, our method does not require hyperparameter tuning on the test set, ensuring better generalization. Furthermore, while they relied on the EU trademark dataset, which included information on color, shape, and categories, our model utilizes a national trademark database enriched with citation networks, VC codes, and ownership details. This approach allows for a deeper exploration of semantic relationships between trademarks, resulting in significantly enhanced performance.

Table 4: Comparison of our model with other works

Approach	AR	NAR
Tursun et al. [7] (TRI-SIFT)	66,117.9	0.07
Tursun et al. [18] (Hand-crafted & CNN)	56,844.1	0.062
Feng et al. [44] (VGG16)	79,538.5	0.086
Feng et al. [44] (SIFT)	79,425.4	0.083
Perez et al. [22] (Visual)	–	0.066
Perez et al. [22] (Conceptual)	–	0.063
Perez et al. [22] (Visual & Conceptual)	–	0.047
Tursun et al. [24] (ATR CAM MAC)	–	0.04
Cao et al. [45] (Attention, Unsupervised)	–	0.051
Tursun et al. [25] (MR-R-SMAC w/URA)	–	0.028
Bernabeu et al. [8] (Weighted fusion: 30% color, 70% shape)	–	0.018
EfficientNet_v2l_Siamese (our model)	**15,498.4**	**0.0169**

Fig. 13 shows the top-10 retrieved results for METU example queries. We select query samples similar to those used in previous models to provide readers with a comparative perspective for evaluation. From

our perspective, we find that our model performs well in retrieving relevant results with varying levels of similarity, and there are fewer completely irrelevant images. However, there are some queries where our model shows certain limitations. First, we observe that the model does not rank the top images as effectively as some other studies. This may be explained by our ACLF, which prioritizes retrieving images with partial similarity rather than attempting to retrieve the most identical ones. Secondly, for queries involving both text and images, retrieval performance is relatively poor. Upon further analysis, this is due to the inadequate separation of text from images. This suggests the need for improvement in text removal models in future studies.

Figure 13: The top-10 retrieved results for METU example queries are displayed in the leftmost column

Through experimental examples, we observe that only about 10% of the images have a similarity score greater than 0. Thus, it may be sufficient to review only around 10% of the returned trademarks, significantly reducing the review time. This still comes with the risk that images with similarity scores less than 0 might still be similar, especially when the text removal process in the image is not optimal. To evaluate this model, we believe that using TPR or TNR would be better than NAR, as NAR can be significantly influenced by very difficult images, overshadowing the evaluation of other images.

5 Conclusions

In this paper, we propose a method to automate the construction of a large labeled image pair dataset by leveraging the rich information available in a national IP Office trademark database. Our study makes several key contributions to the field of TIR. First, we demonstrate the effectiveness of leveraging national trademark databases to automatically generate large labeled datasets, overcoming the limitations of small, manually curated datasets. Second, we introduce an ACLF that better suits the characteristics of trademark data, leading to improved model performance. Third, our experimental results show that the EfficientNet_v2l_Siamese model achieves state-of-the-art performance on the METU dataset, highlighting the potential of our approach. Extensive experiments on various models and parameters reveal that the EfficientNet_v2l model, when integrated into the Siamese architecture and pre-trained on the ImageNet-1k dataset, achieves the best performance. The model achieves the highest performance at both TNR threshold levels, TNR = 0.9 and TNR = 0.95, with TPRs of 77.7% and 70.8% and accuracies of 83.9% and 80.4%. When

evaluated on the METU dataset for comparison with prior studies, our model achieves a new SOTA NAR of 0.0169.

We acknowledge the limitations in our research. Query 5 in Fig. 12 yields suboptimal results, retrieving only 4 out of 10 similar images while returning 5 entirely incorrect results. Notably, the latest model by Tursen et al. [25] also struggled with this query, correctly retrieving only 2 images in the top 10 and 7 in the top 20. We identify that the primary cause of mis-retrieval is the ineffective text removal process. Specifically, the text in the image is not properly removed, leading to feature embeddings being contaminated by textual characteristics. This issue likely arises because the image contains a combination of a white background, red uppercase text positioned prominently on the left side, and graphical elements. Such configurations are rare in the training data used for the text removal model, making it challenging for the model to process effectively. Also, for future research, solutions to reduce input noise caused by incorrect image pair labels should be developed. Additionally, other information fields from the National IP Office trademark data should be explored, such as the historical examination process, to expand the dataset and improve the reliability of input labels. Furthermore, the development and experiment with solutions to enhance the accuracy of text removal models should be implemented, employing more powerful model architectures and utilizing ensemble models to improve the effectiveness of TIR.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Institute of Information Technology, Vietnam Academy of Science and Technology (project number CSCL02.02/22-23) "Research and Development of Methods for Searching Similar Trademark Images Using Machine Learning to Support Trademark Examination in Vietnam".

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Thanh Bui-Minh and Luan Thanh Le; methodology, Thanh Bui-Minh and Nguyen Long Giang; software, Thanh Bui-Minh and Nguyen Long Giang; validation, Thanh Bui-Minh; formal analysis, Thanh Bui-Minh; investigation, Thanh Bui-Minh; resources, Thanh Bui-Minh; data curation, Thanh Bui-Minh; writing—original draft preparation, Thanh Bui-Minh and Luan Thanh Le; writing—review and editing, Thanh Bui-Minh, Nguyen Long Giang and Luan Thanh Le; visualization, Thanh Bui-Minh; supervision, Nguyen Long Giang and Luan Thanh Le; project administration, Thanh Bui-Minh and Luan Thanh Le; funding acquisition, Thanh Bui-Minh. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Castaldi C, Mendonça S. Regions and trademarks: research opportunities and policy insights from leveraging trademarks in regional innovation studies. Reg Stud. 2022;56(2):177–89. doi:10.1080/00343404.2021.2003767.
2. Li X, Yang J, Ma J. Recent developments of content-based image retrieval (CBIR). Neurocomputing. 2021;452:675–89. doi:10.1016/j.neucom.2020.07.139.
3. WIPO. International classification of the figurative elements of marks (Vienna Classification) Ninth Edition [Internet]; 2022. [cited 2024 Feb 6]. Available from: https://www.wipo.int/classifications/vienna/its4vcl/ITSupport_and_download_area/20230101/pdf/vcl_9_en_20230101.pdf.
4. Trappey CV, Trappey AJC, Liu BH. Identify trademark legal case precedents—using machine learning to enable semantic analysis of judgments. World Pat Inf. 2020;62:101980. doi:10.1016/j.wpi.2020.101980.

5. Kucer M, Oyen D, Castorena J, DeepPatent Wu J. Large scale patent drawing recognition and retrieval. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022; 2022; Waikoloa, HI, USA. p. 557–66. doi:10.1109/WACV51458.2022.00063.
6. Liu Y, Ma T. University trademarks: strategies of top Chinese universities. Humanit Soc Sci Commun. 2022;9(1):254. doi:10.1057/s41599-022-01273-7.
7. Tursun O, Kalkan S. METU dataset: a big dataset for benchmarking trademark retrieval. In: Proceedings of the 14th IAPR International Conference on Machine Vision Applications, MVA 2015; 2015; Tokyo, Japan. p. 514–7. doi:10.1109/MVA.2015.7153243.
8. Bernabeu M, Gallego AJ, Pertusa A. Multi-label logo recognition and retrieval based on weighted fusion of neural features. Expert Syst. 2024;41(10):e13627. doi:10.1111/exsy.13627.
9. Duarte V, Zuniga-Jara S, Contreras S. Machine learning and marketing: a systematic literature review. IEEE Access. 2022;10:93273–88. doi:10.1109/ACCESS.2022.3202896.
10. Afifi AJ, Ashour WM. Content-based image retrieval using invariant color and texture features. In: 2012 International Conference on Digital Image Computing Techniques and Applications, DICTA 2012; 2012 Dec 3–5; Fremantle, WA, Australia. doi:10.1109/DICTA.2012.6411665.
11. Le LT. Uncovering import document fraud: leveraging the deep learning approach. Glob Trade Cust J. 2025;20(1):3–10. doi:10.54648/GTCJ2025002.
12. Wu JK. Content-based retrieval for trademark registration. Multimed Tools Appl. 1996;3(3):245–67. doi:10.1007/BF00393940.
13. Jain AK, Vailaya A. Shape-based retrieval: a case study with trademark image databases. Pattern Recognit. 1998;31(9):1369–90. doi:10.1016/S0031-3203(97)00131-3.
14. Lowe DG. Distinctive image features from scale-invariant key points. Int J Comput Vis. 2004;60(2):91–110. doi:10.1023/B:VISI.0000029664.99615.94.
15. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-Up Robust Features (SURF). Comput Vis Image Underst. 2008;110(3):346–59. doi:10.1016/j.cviu.2007.09.014.
16. Ojala T, Pietikäinen M, Mäenpää T. A generalized local binary pattern operator for multiresolution grayscale and rotation invariant texture classification. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Berlin/ Heidelberg, Germany: Springer; 2001. p. 397–406. doi:10.1007/3-540-44732-6_41.
17. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005; 2005; San Diego, CA, USA. p. 886–93. doi:10.1109/CVPR.2005.177.
18. Tursun O, Aker C, Kalkan S. A large-scale dataset and benchmark for similar trademark retrieval. Comput Sci. 2017. doi:10.48550/arXiv.1701.05766.
19. Li X, Abed AM, Shaban M, Le LT, Zhou X, Abdullaev S, et al. Artificial intelligence application for assessment/optimization of a cost-efficient energy system: double-flash geothermal scheme tailored combined heat/power plant. Energy. 2024;313:133594. doi:10.1016/j.energy.2024.133594.
20. Ran L, Yan G, Goyal V, Abdullaev S, Alhomayani FM, Le LT, et al. Advancing solar thermal utilization by optimization of phase change material thermal storage systems: a hybrid approach of artificial neural network (ANN)/genetic algorithm (GA). Case Stud Therm Eng. 2024;64:105513. doi:10.1016/j.csite.2024.105513.
21. Le LT, Xuan-Thi-Thu T. Discovering supply chain operation towards sustainability using machine learning and DES techniques: a case study in Vietnam seafood. Marit Bus Rev. 2024;9(3):243–62. doi:10.1108/MABR-10-2023-0074.
22. Perez CA, Estévez PA, Galdames FJ, Schulz DA, Perez JP, Bastías D, et al. Trademark image retrieval using a combination of deep convolutional neural networks. In: Proceedings of the International Joint Conference on Neural Networks; 2018; Rio de Janeiro, Brazil. doi:10.1109/IJCNN.2018.8489045.
23. USPTO. United States Patent and Trademark Office (USPTO) [Internet] 2024. [cited 2024 Feb 6]. Available from: https://www.uspto.gov/trademark.

24. Tursun O, Denman S, Sivapalan S, Sridharan S, Fookes C, Mau S. Component-based attention for large-scale trademark retrieval. IEEE Trans Inf Forensics Secur. 2022;17:2350–63. doi:10.1109/TIFS.2019.2959921.
25. Tursun O, Denman S, Sridharan S, Fookes C. Learning regional attention over multi-resolution deep convolutional features for trademark retrieval. In: Proceedings of the International Conference on Image Processing, ICIP; 2021; Anchorage, AK, USA. p. 2393–7. doi:10.1109/ICIP42928.2021.9506223.
26. Vesnin D, Levshun D, Chechulin A. Trademark similarity evaluation using a combination of ViT and local features. Information. 2023;14(7):398. doi:10.3390/info14070398.
27. Lim J, Kim S, Park JH, Lee GS, Yang HJ, Lee CW. Recognition of text in wine label images. In: Proceedings of the 2009 Chinese Conference on Pattern Recognition, CCPR 2009, and the 1st CJK Joint Workshop on Pattern Recognition, CJKPR; 2009; Nanjing, China. p. 911–15. doi:10.1109/CCPR.2009.5343972.
28. Li X, Yang J, Ma J. CNN-SIFT consecutive searching and matching for wine label retrieval. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Cham, Switzerland: Springer; 2019. p. 250–61. doi:10.1007/978-3-030-26763-6_24.
29. Li X, Yang J, Ma J. Large scale category-structured image retrieval for object identification through supervised learning of CNN and SURF-based matching. IEEE Access. 2020;8:57796–809. doi:10.1109/ACCESS.2020.2982560.
30. Li Q, Ma L, Jiang Z, Li M, Jin B. TECMH: transformer-based cross-modal hashing for fine-grained image-text retrieval. Comput Mater Contin. 2023;75(2):3713–28. doi:10.32604/cmc.2023.037463.
31. Waqas U, Visser JW, Choe H, Lee D. Multimodal fused deep learning networks for domain specific image similarity search. Comput Mater Contin. 2023;75(1):243–58. doi:10.32604/cmc.2023.035716.
32. Trappey AJC, Trappey CV, Shih S. An intelligent content-based image retrieval methodology using transfer learning for digital IP protection. Adv Eng Inform. 2021;48:101291. doi:10.1016/j.aei.2021.101291.
33. Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2015; Boston, MA, USA. p. 4353–61. doi:10.1109/CVPR.2015.7299064.
34. Khayyat MM, Elrefaei LA. Manuscripts image retrieval using deep learning incorporating a variety of fusion levels. IEEE Access. 2020;8:136460–86. doi:10.1109/ACCESS.2020.3010882.
35. Yan S, Qi Y, Liu M, Wang Y, Liu B. Object tracking based on siamese network with 3D attention and multiple graph attention. Comput Vis Image Underst. 2023;235:103786. doi:10.1016/j.cviu.2023.103786.
36. Lan T, Feng X, Li L, Xia Z. Similar trademark image retrieval based on convolutional neural network and constraint theory. In: Proceedings of the 2018 8th International Conference on Image Processing Theory, Tools and Applications, IPTA 2018; 2018 Nov 07–10; Xi'an, China. doi:10.1109/IPTA.2018.8608162.
37. Trappey CV, Trappey AJC, Lin SCC. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. Adv Eng Inform. 2020;45:101120. doi:10.1016/j.aei.2020.101120.
38. Tursun O, Denman S, Sridharan S, Fookes C. Learning test-time augmentation for content-based image retrieval. Comput Vis Image Underst. 2022;222:103494. doi:10.1016/j.cviu.2022.103494.
39. IPVietnam. Intellectual property office of Vietnam [Internet]. 2024. [cited 2024 Feb 6]. Available from: http://wipopublish.ipvietnam.gov.vn.
40. Kumar GVRM, Madhavi D. Stacked Siamese Neural Network (SSiNN) on neural codes for content-based image retrieval. IEEE Access. 2023;11:77452–63. doi:10.1109/ACCESS.2023.3298216.
41. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2006; New York, NY, USA. p. 1735–42. doi:10.1109/CVPR.2006.100.
42. Fan L, Sun X, Rosin PL. Attention-modulated triplet network for face sketch recognition. IEEE Access. 2021;9:12914–21. doi:10.1109/ACCESS.2021.3049639.
43. Gatis D. Rembg [Internet]. [cited 2024 Feb 6]. Available from: https://github.com/danielgatis/rembg.
44. Feng Y, Shi C, Qi C, Xu J, Xiao B, Wang C. Aggregation of reversal invariant features from edge images for large-scale trademark retrieval. In: Proceedings of the 2018 4th International Conference on Control, Automation and Robotics, ICCAR 2018; 2018; Auckland, New Zealand. p. 384–8. doi:10.1109/ICCAR.2018.8384705.
45. Cao J, Huang Y, Dai Q, Ling WK. Unsupervised trademark retrieval method based on attention mechanism. Sensors. 2021;21(5):1–18. doi:10.3390/s21051894.