©2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This a post-print of the article “Content-based Image Retrieval using an Adaptive Triplet Learning Approach from Multi-label Remote Sensing Image Archives” published in the Proceedings of ICARES 2025 - 2025 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology.

The published article is available on https://doi.org/10.1109/ICARES67579.2025.11371530

A. K R, A. Preethy Byju and A. S. Garea, "Content-based Image Retrieval using an Adaptive Triplet Learning Approach from Multi-label Remote Sensing Image Archives," 2025 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES), Surabaya, Indonesia, 2025, pp. 1-7, doi: 10.1109/ICARES67579.2025.11371530.

Content-based Image Retrieval using an Adaptive Triplet Learning Approach from Multi-label Remote Sensing Image Archives

1st Aswathy K R
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Amritapuri
Kollam, India
aswathykr@am.amrita.edu

2nd Akshara Preethy Byju
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Amritapuri
Kollam, India
aksharapb@am.amrita.edu

3rd Alberto S. Garea
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
Santiago de Compostela, Spain
jorge.suarez.garea@usc.es

***Abstract*—Rapid expansion of remote sensing (RS) image archives necessitates the development of advanced content-based image retrieval (CBIR) methods to effectively organize and access high-dimensional visual data. Deep metric learning, particularly with triplet loss, has shown promising results in retrieval tasks by embedding semantically similar images closer in feature space. However, conventional fixed-margin triplet loss functions are inadequate for multi-label RS archives, where high intra-class variability and inter-class similarity is observed. These limitations often result in ineffective training due to the dominance of trivial triplets and the zero-loss problem. In this work, we propose a novel multi-label adaptive triplet learning approach for CBIR in RS archives. The proposed method introduces an adaptive margin approach that dynamically adjusts the triplet loss margin based on the hardness of each sample, computed as the difference between anchor-positive and anchor-negative distances. This adaptive strategy enables the model to focus on informative (hard) triplets and suppresses the influence of trivial ones, leading to more discriminative feature embeddings. We evaluate our proposed approach, on two challenging multi-label RS archives: MLRSNet and UCMerced. Experimental results demonstrate that the proposed method consistently outperforms fixed-margin triplet loss and several state-of-the-art CBIR approaches, particularly in retrieval accuracy and robustness. These findings confirm the effectiveness of adaptive margin learning in handling the complexities of multi-label RS image retrieval.**

***Index Terms*—Content-based image retrieval (CBIR), remote sensing, deep metric learning, multi-label image retrieval, triplet loss, adaptive margin.**

I. Introduction

Advancements in remote sensing (RS) technology have led to the accumulation of vast amounts of high-resolution imagery. Efficient management and retrieval of relevant contents from such massive archives is a challenging task to perform. Traditionally, content-based image retrieval (CBIR) methods in RS have relied on hand-crafted features obtained from color histograms, texture descriptors (e.g., GLCM), and local descriptors like scale-invariant feature transform (SIFT) and local binary pattern (LBP) [1] [2]. Although these conventional methods showcased outstanding performance, they often failed to capture the complex semantic information inherent within the RS images, especially under variations in viewpoint, scale, and illumination.

With the rise of deep learning, CNN-based methods have shown superior performance in image representation and retrieval tasks. In particular, deep metric learning that considers triplet loss has demonstrated efficient performance, as it allows learning of an embedding space where semantically similar images are closer together and dissimilar ones are farther apart [3], [4], [5]. The traditional triplet loss function trains on image triplets (anchor, positive, negative), aiming to reduce the distance between the anchor and positive images while increasing the distance to the negative image by at least a margin [6].

However, a key limitation of the traditional triplet loss is its use of a fixed margin to separate positive and negative pairs. This is challenging when considering multi-label RS image archives, where images often share some semantic content. For example, an image labeled with both ”river” and ”forest” may still be semantically similar to an image labeled only ”forest.” In such cases, a fixed margin may lead to over-penalizing meaningful similarities or under-emphasizing subtle differences [7]. Additionally, fixed-margin triplet loss suffers from the zero-loss problem, where several triplets become non-informative (i.e., already satisfy the margin condition), leading to slow convergence and limited feature discrimination [8]. These issues are particularly pronounced in RS imagery, where high intra-class variability and inter-class similarity are common.

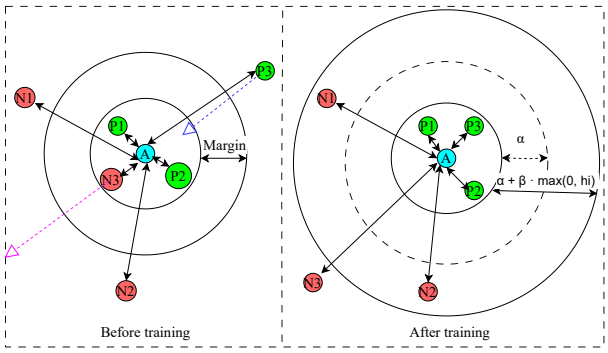


Fig. 1. Triplet training using adaptive margin triplet loss highlighting the margin change for hard triplets. (A, P_1, N_1) and (A, P_2, N_2) are trivial triplets whereas (A, P_3, N_3) is a hard triplet. After trained using our adaptive margin triplet training approach, negative image $N3$ will be pushed away from the anchor A and positive image $P3$ will be pulled closer to anchor image A.

To overcome these limitations, we propose a novel multi-label adaptive triplet learning approach to perform CBIR from RS image archives. Instead of a static margin, we employ an adaptive margin based on the relative hardness of each triplet—measured by the difference between the distances of anchor-positive and anchor-negative pairs (see Fig. 1). This encourages the model to focus on hard and informative examples while ignoring trivial ones. By dynamically adjusting the decision boundary, our method is better suited for multi-label retrieval scenarios. We evaluate our proposed approach on two multi-label RS benchmark archives: MLRSNet [9] and UCMerced [10] land use archive. Our approach consistently outperforms fixed-margin triplet loss and other baseline retrieval methods, demonstrating its effectiveness for learning robust embeddings in multi-label RS settings. The key contributions of our work are three-fold:

- We propose a novel adaptive-margin triplet learning framework tailored for multi-label CBIR in RS image archives.
- We propose a hardness-aware margin adjustment strategy to improve retrieval performance and learning efficiency.
- We validate our method on two benchmark RS archives, showing significant improvements over traditional metric learning approaches.

The remainder of this paper is organized as: Section II provides the state-of-the-art of the proposed work, Section III details about the novel triplet learning approach proposed, Section IV provides the description of the experiments, Section V presents the experimental results and finallt, section VI concludes the work.

II. Related Work

Triplet-based deep metric learning has become a widely adopted framework for content-based image retrieval (CBIR), enabling the learning of an embedding space where semantically similar images are positioned closer together while dissimilar ones are pushed farther apart. Earlier works such as Siamese networks proposed by Bromley et al. [11] laid the foundation for metric learning, which was later extended by Schroff et al. [4] introduced the triplet loss in the FaceNet model for face recognition, while Hermans et al. [5] demonstrated its effectiveness for person re-identification tasks.

In RS, triplet networks have been explored for RS image retrieval. Hoffer and Ailon [12] applied triplet learning for visual similarity, and Cao et al. [13] proposed a deep metric learning framework tailored for RS archives,while Demir and Bruzzone [1]introduced an active learning method for relevance feedback in RS CBIR. Sumbul et al. [14] introduced the DAS-RHDIS model that selects informative triplets for multi-label RS retrieval. These approaches demonstrated the potential of triplet loss in RS CBIR, but primarily rely on a fixed margin formulation.Furthermore, the introduction of large-scale multi-label archives such as BigEarthNet [15] has emphasized the need for retrieval methods capable of handling diverse and co-occurring land-cover categories in realistic remote sensing scenarios.

However, fixed-margin triplet loss suffers from the zero-loss problem, where many triplets contribute no learning once they satisfy the margin constraint thereby slowing down training [7], [8]. This is particularly problematic in multi-label RS image archives, where partial label overlaps are common. For instance, an image labeled with ['buildings', 'cars'] may be semantically close to one labeled only with ['cars'], which makes fixed-margin decisions ambiguous. In such scenarios, using a static margin can misrepresent the fine-grained semantic relationships among images.

To address these limitations, Ha and Blanz [16] proposed an adaptive margin triplet loss for image ranking tasks. Their approach adjusts the margin for each triplet based on perceptual differences, improving training stability and eliminating the need for complex triplet mining strategies. While their focus was on aesthetic and similarity rating archives, the core concept of dynamically adapting margins supports our motivation.

Building on these ideas, our proposed method introduces a hardness-aware adaptive margin strategy specifically designed for multi-label RS image retrieval. Unlike prior methods such as TNDML [12], RSDML [13], and DAS-RHDIS [14], our approach adapts the margin based on triplet hardness, enabling the model to focus on informative examples while ignoring trivial ones, leading to improved feature discrimination in complex multi-label RS scenarios.

III. METHODOLOGY

A. Problem Formulation

Let $X = \{X_i\}_{i=1}^{M}$ denote a set of M RS images, where each image X_i is associated with a set of multiple labels $Y_i \subseteq \mathcal{C}$, where $\mathcal{C}$ is the set of possible labels from the image archive. Given a query image X_q, our objective is to retrieve the top-k semantically similar images from the archive based on their similar semantic characteristics. We propose a novel triplet-based training framework, where each training unit consists of a triplet (A, P, N): an *anchor* image A, a *positive* image

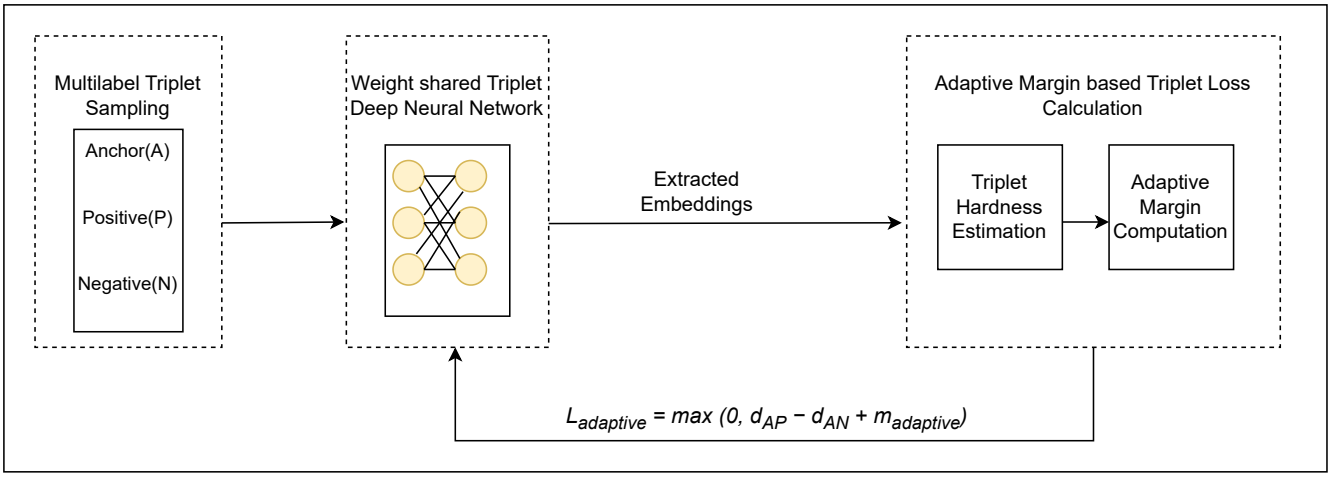


Fig. 2. Block diagram of the proposed multi-label adaptive margin triplet training framework.

P (with overlapping labels), and a *negative* image N (with minimal or no label overlap). The core of our method is an adaptive margin triplet loss that dynamically adjusts based on triplet hardness, thus improving discriminative embedding learning for multi-label image retrieval. The architecture of the proposed framework is illustrated in Fig. 2, which focuses on the triplet-based learning process with an adaptive margin loss. The proposed framework integrates mainly three modules: i) multi-label triplet sampling; ii) shared weight triplet deep neural network and iii) computation of an adaptive margin triplet loss which obtains a triplet hardness estimation and adaptive margin computation. These modules are detailed in the following sub-sections below.

B. Multi-label Triplet Sampling

In multi-label archives, each image X_i is associated with a set of labels $Y_i \subseteq \mathcal{C}$, where $\mathcal{C}$ represents the set of possible classes. To construct effective triplets for metric learning, we employ label-based similarity approach to select appropriate positive and negative samples. Given an anchor image A with label set Y_A, we define semantic similarity between two images X_i and X_j based on their label sets Y_i and Y_j using the Jaccard similarity coefficient as:

$$\text{Sim}(X_i, X_j) = \frac{|Y_i \cap Y_j|}{|Y_i \cup Y_j|} \tag{1}$$

A positive image P is selected such that its label similarity with the anchor satisfies the constraint as:

$$\text{Sim}(A, P) = \frac{|Y_A \cap Y_P|}{|Y_A \cup Y_P|} \geq \tau_p \tag{2}$$

ensuring that the anchor and positive share a sufficient number of labels. Similarly, a negative image N is selected such that:

$$\text{Sim}(A, N) = \frac{|Y_A \cap Y_N|}{|Y_A \cup Y_N|} \leq \tau_n \tag{3}$$

indicating low or no label similarity with the anchor. Here, τ_p and τ_n are pre-defined similarity thresholds that determine the minimum required semantic overlap for a positive pair and the maximum allowable overlap for a negative pair, respectively. In our experiments, we set $\tau_p = 0.5$ and $\tau_n = 0$, as detailed in Section IV. This sampling strategy ensures that positive and negative examples are selected based on actual semantic overlap, making the triplet structure suitable for multi-label image retrieval tasks.

C. Triplet Hardness Estimation

Given a sampled triplet (A, P, N), we extract deep embeddings f_A, f_P, and f_N using an appropriate CNN backbone network. We then compute the Euclidean distances between Anchor-Positive and Anchor-Negative image as:

$$d_{AP} = \|f_A - f_P\|_2, \quad d_{AN} = \|f_A - f_N\|_2 \tag{4}$$

Here, d_{AP} represents the distance between the anchor and the positive sample in the embedding space, and d_{AN} represents the distance between the anchor and the negative sample. These distances quantify how similar or dissimilar the samples are in the learned feature space. The *triplet hardness (h)* is defined as:

$$h = d_{AP} - d_{AN} \tag{5}$$

A high positive h value indicates a hard triplet, where the negative is closer to the anchor than the positive.

D. Adaptive Margin Computation

To effectively weight triplets based on their hardness, we propose a dynamic margin as:

$$m_{\text{adaptive}} = \alpha + \beta \cdot \max(0, h) \tag{6}$$

where α is a base margin and β is a scaling factor that determines how much the margin increases for harder triplets.

E. Adaptive Triplet Loss

The proposed loss function incorporates the adaptive margin as:

$$\mathcal{L}_{\text{adaptive}} = \max\left(0, d_{AP} - d_{AN} + m_{\text{adaptive}}\right) \tag{7}$$

This encourages the model to focus on hard and informative triplets while suppressing trivial ones, effectively mitigating the zero-loss issue in conventional fixed-margin approaches.

F. Retrieval Process

Once the model is trained, semantic embeddings are extracted for both query and test images. Retrieval is performed by computing the cosine similarity between the query embedding (f_q) and all test embeddings (f_i):

$$\text{Simcos}(f_q, f_i) = \frac{f_q \cdot f_i}{\|f_q\|\|f_i\|}$$

Where:

- f_q = embedding of the query image.
- f_i = embedding of the i-th test image.
- $f_q \cdot f_i$ = dot product between the query and test image embeddings.
- $\|f_q\|$ = L2 norm (Euclidean norm) of the query embedding.
- $\|f_i\|$ = L2 norm of the test image embedding.

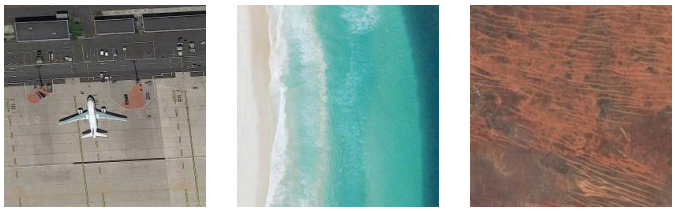
Cosine similarity measures the cosine of the angle between the two embeddings. A value close to 1 indicates high similarity, 0 indicates no similarity, and -1 indicates complete dissimilarity.

IV. Dataset Description and Design of Experiments

A. Dataset Description

To evaluate the performance of the proposed novel approach to multi-label adaptive triplet learning for RS image retrieval, we conducted experiments on two benchmark multi-label RS archives: a selected subset of the MLRSNet [9] archive and a multi-label annotated version of the UCMerced [10] Land Use archive. The MLRSNet archive is a large-scale RS image archive originally consisting of over 100,000 images labeled with up to 13 labels per image across 60 categories. The labels in MLRSNet were generated by combining existing land-cover maps and GIS information with high-resolution Google Earth and Bing Maps imagery, followed by manual verification by experts to ensure accuracy [9].For our experiments, we selected a subset containing 15 representative classes: airplane, bare-soil, buildings, cars, chaparral, court, dock, field, grass, mobile-home, pavement, sand, sea, ship, and tanks. The selected subset contains 30,000 images, each resized to 256×256 pixels. These images are collected from Google Earth and Bing Maps and represent diverse land-cover types with high intra-class variability and frequent label co-occurrence, making the archive well-suited for evaluating multi-label CBIR systems. The spatial resolution of the images varies, but typically each pixel corresponds to a few meters on the ground, depending on the altitude and sensor used during the image acquisition. Fig.3 shows an example of images from the MLRSNet with the associated multilabel annotations.

The second archive is a multi-label version of the UCMerced Land Use [10] archive, which originally contains 2,100 aerial images evenly distributed among 21 land-use scene classes. In this study, we use a modified version of the archive [17] where each image is annotated with one or more semantic labels, resulting in a total of 17 distinct labels: airplane, bare-soil, buildings, cars, chaparral, court, dock, field, grass, mobile-home, pavement, sand, sea, ship, tanks, trees, and water. Each image in this multi-label variant typically contains between 1 to 5 labels.The annotations were created through manual re-labeling by domain experts,extending the original single-label UCMerced dataset into a multi-label archive,thereby better capturing object co-occurrence within a scene The image resolution is 256×256 pixels, and the spatial resolution of the original aerial images from the UCMerced archive is approximately 0.3 meters per pixel. This enriched label set introduces increased complexity due to object co-occurrence within a scene, allowing for a more realistic evaluation of multi-label CBIR performance. Fig.4 shows an example of images from the UCMerced with the associated multilabel annotations.



airplane, airport, buildings, cars, pavement, trees | *beach, sand, sea, water* | *desert, sand, chaparral*

Fig. 3. Examples of images from the MLRSNet archive and the multilabels associated with them.

dock, ship, water | *grass, sand, trees* | *bare soil, cars, court, pavement, trees*

Fig. 4. Examples of images from the UCMerced Land Use archive and the multilabels associated with them.

B. Design of Experiments and Sensitivity Analysis

To evaluate the effectiveness of the proposed novel approach to multi-label adaptive triplet learning for RS image retrieval model, we considered two CNN architectures as backbone networks: ResNet18 [18] and DenseNet121 [19]. Data augmentation techniques, including random cropping and

TABLE I
FORMULAS FOR THE EVALUATION METRICS USED IN THIS WORK, INCLUDING ACCURACY, PRECISION, RECALL, F1-SCORE, AND ANMRR

Performance metric	Formula
Accuracy	$\frac{1}{\lvert X^{\text{final}}\rvert}\sum_{r=1}^{\lvert X^{\text{final}}\rvert}\frac{\lvert Y_q\cap Y_r\rvert}{\lvert Y_q\cup Y_r\rvert}$
Precision	$\frac{1}{\lvert X^{\text{final}}\rvert}\sum_{r=1}^{\lvert X^{\text{final}}\rvert}\frac{\lvert Y_q\cap Y_r\rvert}{\lvert Y_r\rvert}$
Recall	$\frac{1}{\lvert X^{\text{final}}\rvert}\sum_{r=1}^{\lvert X^{\text{final}}\rvert}\frac{\lvert Y_q\cap Y_r\rvert}{\lvert Y_q\rvert}$
F1-score	$\frac{1}{\lvert X^{\text{final}}\rvert}\sum_{r=1}^{\lvert X^{\text{final}}\rvert}\frac{2\cdot\text{Precision}_r\cdot\text{Recall}_r}{\text{Precision}_r+\text{Recall}_r}$
ANMRR	$\frac{1}{Q}\sum_{q=1}^{Q}\frac{\frac{1}{N_q}\sum_{j=1}^{N_q}\min(R_{q,j},K+1)-\frac{N_q+1}{2}}{K-\frac{N_q+1}{2}}$

horizontal flipping, were applied during training to improve the model's generalization. The optimal learning rate was found to be 0.0001, and the batch size was set to 32. All experiments were performed on an NVIDIA P100 GPU. The archives used for training were split into training (80%), validation (10%), and test (10%) sets.

For quantitative evaluation, we used the following metrics: Accuracy, F1-score, Recall, Precision, and Average Normalized Modified Retrieval Rank (ANMRR). For each query image selected from the test set, the top 15 most similar images were retrieved based on cosine similarity in the embedding space. Precision, Recall, and F1-score are computed using sample-based averaging as implemented in scikit-learn [20], while ANMRR follows the retrieval evaluation formulation widely adopted in recent remote sensing studies [21]. The mathematical definitions of these metrics are summarized in Table I. Here, X^{final} denotes the set of all test queries, Y_q is the label set of query image X_q, and Y_r is the label set of the retrieved image X_r. For ANMRR, $R_{q,j}$ is the rank of the j-th relevant image for query X_q, N_q is the number of relevant images, K is the cutoff rank, and Q is the total number of queries.

To implement the triplet loss effectively, we used a label-based similarity measure defined in Equation 1. For the triplet selection, a positive image was defined based on the condition in Equation 2, where we set the similarity threshold to $\tau_p = 0.5$, ensuring that the anchor and positive share a sufficient number of common labels. Conversely, a negative image was selected according to Equation 3, with a threshold $\tau_n = 0$, indicating low or no label similarity with the anchor.

We conducted a sensitivity analysis on the adaptive margin parameters α and β using ResNet18 and DenseNet121 as backbone networks, evaluated on the UCMerced and MLRSNet archives. The experiments were performed with α values of 0.3, 0.5, and 0.7, and β values of 0.1, 0.2, and 0.3. For each α–β combination, we plotted the validation loss against training epochs to assess the effect of the margin parameters on model convergence.

For ResNet18 trained on the UCMerced archive, the optimal performance was observed with $\alpha = 0.7$ and $\beta = 0.1$, showing the lowest validation loss and better convergence behavior. Figure 5 illustrates these results, highlighting the influence of the adaptive margin parameters on training stability and retrieval performance.

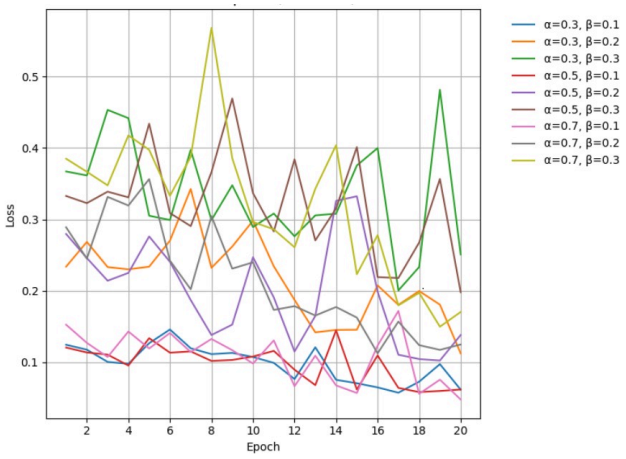


Fig. 5. Optimal values for α and β for ResNet18 network using UCMerced.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we evaluate the performance of the proposed multi-label adaptive triplet learning model on both the *M*LRSNet and *U*CMerced Land Use archives using quantitative metrics and qualitative retrieval examples.

We first investigate the impact of different backbone networks on the performance of our approach. To this end, we conduct experiments using two widely used deep learning architectures: ResNet18 [18] and DenseNet121 [19]. The performance metrics obtained for both archives are summarized in Table II. It can be observed that DenseNet121 performs slightly better on the MLRSNet dataset, while ResNet18 achieves a clear performance gain on the UCMerced archive across all evaluation metrics. This indicates that the effectiveness of the backbone network may vary with dataset characteristics; however, ResNet18 shows better generalization and consistency, and is therefore adopted as the backbone network for all subsequent experiments.

We further compare our proposed multi-label adaptive triplet learning approach with existing state-of-the-art methods including TNDML [12], RSDML [13], and DAS-RHDIS [14]. Among these, all three approaches employ triplet-based deep metric learning frameworks using a fixed-margin triplet loss strategy. In contrast, our approach introduces an adaptive-margin formulation, which dynamically adjusts the margin based on triplet hardness. This adaptive mechanism provides greater flexibility in handling the semantic overlap present in multi-label RS image archives. Table III shows that the proposed method achieves superior performance compared to all baseline models on the UCMerced archive. These results demonstrate the effectiveness of adaptive margin formulation in enhancing retrieval performance over conventional fixed-

TABLE II
PERFORMANCE METRICS WITH VARYING BACKBONE NETWORKS IN PROPOSED METHOD.

Archive	Backbone Network	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ANMRR (%)
MLRSNet	**DenseNet121**	**69.8**	**76.1**	**78.1**	**75.9**	**0.21**
	ResNet18	67.6	77.7	75.5	74.9	0.21
UCMerced	DenseNet121	56.2	60.9	65.3	61.4	0.33
	ResNet18	**74.5**	**77.8**	**82.0**	**78.7**	**0.27**

TABLE III
COMPARISON OF OUR PROPOSED APPROACH USING RESNET18 BACKBONE NETWORK WITH STATE-OF-THE-ART METHODS ON THE UCMERCED ARCHIVE.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
TNDML [12]	44.0	52.6	55.8	54.2
RSDML [13]	48.4	56.3	61.9	59.0
DAS-RHDIS [14]	56.8	65.3	70	67.5
Proposed Approach	**74.5**	**77.8**	**82.0**	**78.7**

margin approaches. Thus, we can observe that the proposed approach considers the rich semantic embeddings found in the label information that is utilized to improve the retrieval performance.

Qualitative retrieval results are illustrated in Figure 6, showcasing the top retrieved images for randomly selected queries from the UCMerced archive. As observed, our proposed method retrieves semantically relevant images even in the presence of intra-class variation and inter-class similarity. This highlights the robustness and discriminative power of the learned embeddings, further confirming the utility of the adaptive margin strategy in RS image retrieval tasks.

Fig. 6. Qualitative results of the proposed approach using ResNet18 trained on the UCMerced archive. The 1^{st}, 10^{th}, and 15^{th} retrieved images are shown with their predicted labels.

VI. CONCLUSION

In this study, we propose a novel multi-label adaptive triplet learning approach for CBIR in RS image archive. The proposed method introduces an adaptive margin-based triplet training strategy that adjusts the margin dynamically, addressing the limitations of traditional fixed-margin triplet loss by ensuring hard triplets are given higher importance during the learning process. We evaluated the proposed method on two benchmark archives, UCMerced and MLRSNet, using DenseNet121 and ResNet18 as the backbone networks. The experimental results demonstrate that the adaptive margin approach significantly improves retrieval performance, enhancing the discriminative power of feature embeddings. This method effectively handles the challenges of multi-label RS image retrieval, where intra-class variation and inter-class similarity can be high. Overall, our proposed method outperforms existing state-of-the-art methods in both accuracy and retrieval performance, making it a promising solution for CBIR in RS applications.

REFERENCES

[1] M. Dateu and K. Seidel, "Image information mining: exploration of earth observation archives," *Geographica Helvetica*, vol. 58, no. 2, pp. 154–168, 2003.

[2] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for sar images," *IEEE transactions on geoscience and remote sensing*, vol. 54, no. 8, pp. 4806–4817, 2016.

[3] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[5] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[6] Y. Feng, H. Wang, H. R. Hu, L. Yu, W. Wang, and S. Wang, "Triplet distillation for deep face recognition," in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 808–812, IEEE, 2020.

[7] W. Ge, "Deep metric learning with hierarchical triplet loss," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 269–285, 2018.

[8] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1386–1393, 2014.

[9] X. Qi, P. Zhu, Y. Wang, L. Zhang, J. Peng, M. Wu, J. Chen, X. Zhao, N. Zang, and P. T. Mathiopoulos, "Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 337–350, 2020.

[10] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

[11] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 737–744, 1994.

[12] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pp. 84–92, Springer, 2015.

[13] R. Cao, Q. Zhang, J. Zhu, Q. Li, Q. Li, B. Liu, and G. Qiu, "Enhancing remote sensing image retrieval using a triplet deep metric learning network," *International Journal of Remote Sensing*, vol. 41, no. 2, pp. 740–751, 2020.

[14] G. Sumbul, M. Ravanbakhsh, and B. Demir, "Informative and representative triplet selection for multilabel remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[15] H. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5901–5904, IEEE, 2019.

[16] M. L. Ha and V. Blanz, "Deep ranking with adaptive margin triplet loss," *arXiv preprint arXiv:2107.06187*, 2021.

[17] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] H. Sumbul, L. Luppino, and B. Demir, "Informative instance-based image retrieval in remote sensing archives," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.