IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 April 2023, accepted 14 April 2023, date of publication 17 April 2023, date of current version 21 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3267804

APPLIED RESEARCH

Universal Image Embedding: Retaining and Expanding Knowledge With Multi-Domain Fine-Tuning

SOCRATIS GKELIOS[1], ANESTIS KASTELLOS[2], YIANNIS S. BOUTALIS[1], (Senior Member, IEEE), AND SAVVAS A. CHATZICHRISTOFIS[2], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, Democritus University of Thrace, Kimmeria, 67100 Xanthi, Greece
[2]Department of Computer Science, Intelligent Systems Laboratory, Neapolis University Pafos, 8042 Paphos, Cyprus

Corresponding author: Savvas A. Chatzichristofis (s.chatzichristofis@nup.ac.cy)

This work was supported in part by the Democritus University of Thrace; in part by Neapolis Academic Enterprises Ltd.; and in part by the European Union and Greek National Funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, by the Call RESEARCH-CREATE-INNOVATE, under Project T2EDK-02743.

ABSTRACT The overall purpose of this study is to propose a novel fine-tuning method for the CLIP architecture that enables the retention of pre-existing knowledge from large datasets and the creation of a domain-agnostic image encoder for universal image embedding, addressing the challenge of transferring knowledge from source to target tasks using deep learning models. The basic design of the study involves applying the proposed method directly (without fine-tuning) to a wide range of instance retrieval and recognition tasks to evaluate its effectiveness. The study's major findings indicate that the proposed method significantly enhances performance on unseen domains without requiring separate fine-tuning for each domain. The authors' success in the Google Universal Image Embedding competition, where they were awarded a Gold medal out of 1200 teams, inspired their proposed method. These results have significant implications for real-life applications where multiple domains are common. In conclusion, the study offers a practical solution for transfer learning that addresses the challenges of dealing with multiple domains and advances deep learning, potentially inspiring further research in this area and driving progress in the field.

INDEX TERMS CBIR, vision transformers, CLIP, deep learning, global features, image retrieval, universal image embedding, local features.

I. INTRODUCTION

The effectiveness of deep learning models in generalizing their knowledge to new scenarios has been a matter of significant interest in the machine learning community, particularly in light of the proliferation of both vision and language models and the ongoing expansion of available datasets. Pre-trained vision models trained on large datasets, such as Imagenet [1], often perform well on various computer vision tasks [2], [3] [4], [5].As the size and complexity of these models grow, techniques such as distillation [6], [7] and federated learning [8] have emerged as promising approaches to handle bigger models efficiently while maintaining their adaptability to various tasks and data distributions.However, they can struggle when the data distribution between the training dataset and the target task differs significantly [9]. For example, a pre-trained ResNet architecture trained on ImageNet may perform poorly when applied to instance recognition tasks. To achieve satisfactory performance, fine-tuning such models using datasets more closely related to the target task is often necessary.

Various techniques have been suggested for addressing instance recognition problems by incorporating intricate loss functions to enable the model to distinguish between different instances that may occur. While these approaches have demonstrated promising outcomes, they necessitate the

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

fine-tuning of the model for each benchmark individually, as training a single model to recognize a diverse range of objects can be challenging due to the distinctive characteristics and features inherent to each object domain. For instance, a model trained to recognize landmarks may need to learn to identify specific architectural elements, while a model trained to recognize products may need to recognize branding logos or packaging. As a result, it is common for researchers to focus on a single object domain when developing recognition models, allowing them to create models specialized for that particular domain and achieve high accuracy. However, this process is time-consuming, as it requires the creation of a separate model for each individual object domain, and is also not reflective of real-life applications, which are not restricted to a single domain.

The Google Universal Image Embedding competition hosted on Kaggle[1] aimed to assess the universality and generalization ability of various techniques through the presentation of three main challenges: (1) the absence of a provided training dataset, requiring competitors to assemble their own datasets for model training, (2) the significant differences in categories in the test distribution, requiring the model to handle intra-class similarity across a wide range of categories with dissimilar characteristics in an image retrieval task at the instance level, and (3) a constraint on the size of the embedding to a maximum of 64.

Our team achieved 6th place in the Google Universal Image Embedding competition. Our final solution was based on the CLIP [10] architecture and involved training on the LAION-2B dataset, a subset of LAION-5B [11]. We also used the ArcFace [12] loss function, a popular loss function in Kaggle competitions that increases intra-class separability, specifically employing the SubCenter ArcFace [13] variant of the original ArcFace with dynamic margins [14].

Inspired by the Google Universal Image Embedding competition, our goal is to further evaluate the performance of our fine-tuned model on a range of challenging instance recognition and retrieval benchmarks.The motivation behind this research is to develop a domain-agnostic image encoder that can perform well across different domains without the need for individual fine-tuning for each domain. Also, the multi-domain fine-tuning field for instance retrieval is novel, and there is limited prior work in this area, if any. In this regard, we share some architectural choices that were refined during the competition to maximize performance. We thoroughly assess the model's discriminatory capability by utilizing it as a black box image descriptor extractor in both general and unseen domains (during the training phase). Given a pre-trained neural network f with parameters θ and a database of N images $I_1, I_2, \ldots, I_N$, the goal of fine-tuning is to optimize the parameters θ of f to better fit the image database.

[1]https://www.kaggle.com/competitions/google-universal-image-embedding

The optimization problem can be formulated as:

$$\theta^* = argmin_\theta L(I, y; \theta) \quad (1)$$

where L is a loss function that measures the difference between the predicted output y and the true label of image I, and θ^* is the optimized set of parameters that minimize the loss function.

When the training set distribution is significantly different from the fine-tuning set, the model might suffer from poor generalization, leading to a suboptimal performance on the target task. In contrast, when the fine-tuning set and the test set share similar distributions, the model is more likely to adapt effectively, reducing the loss function value and improving its performance on the target task.

Let's consider two fine-tuning sets, A and B, where A has a similar distribution to the test set, and B has a significantly different distribution. After fine-tuning the model on both sets, we obtain the optimized parameters θ_A^* and θ_B^*. The loss function values for these two sets can be represented as L_A and L_B, respectively. Since the fine-tuning set A has a similar distribution to the test set, the loss function value L_A is expected to be lower than L_B. In general, when the test set distribution differs significantly from the training set, it becomes challenging to optimize the model for high performance on the test set, as it contains previously unseen data samples.

Once the neural network is fine-tuned, the goal of image retrieval is to find the image I_i in the database that is most similar to the query image I_q. This can be mathematically expressed as:

$$I_i = argmax_{sim}(f(I_q; \theta^*), f(I_j; \theta^*)), \quad j = 1, 2, \ldots, N \quad (2)$$

where $f(I; \theta^*)$ is the fine-tuned neural network with optimized parameters $\theta*$, and *sim* is a similarity function that measures the similarity between the feature representations of the query image and the database image.

As the loss function value L_A is lower than L_B, the model fine-tuned on set A (with parameters θ_A) is expected to yield better image retrieval results than the model fine-tuned on set B (with parameters θ_B), as the similarity function will produce higher values for the model adapted to the similar distribution.

In summary, the process of image retrieval using pre-trained and fine-tuned neural networks involves optimizing the parameters of a pre-trained neural network using a loss function and then using the fine-tuned network to compute the similarity between the query image and the database images. In our case, our contribution can be summarized as follows:

- The development of a domain-agnostic image encoder that emphasizes the universality of the embeddings.
- A novel training scheme that involves freezing specific parts of the model, using distinct learning rates for the head and backbone, and exposing the model to instance categorization tasks involving multiple domains.
- Demonstration of the superiority of our fine-tuned model to the pre-trained CLIP on all benchmarks,

with exceptional results and even state-of-the-art performance in some cases.

The structure of this paper is as follows. Section II provides an overview of related work that is directly relevant to this paper. Section III discusses the system architecture, pilot employed in this study, and outlines the proposed scenario. In Section IV, the authors assess the proposed scenario and present the results obtained. Finally, Section V summarizes the conclusions drawn from this study.

II. RELATED WORK

A. IMAGE-TEXT FOUNDATION MODELS

Image-text foundation models, which combine visual information with raw text descriptions for classification tasks, have gained popularity in recent years as an alternative to traditional visual-based classification models. These models do not have the restriction of a predetermined number of classes, which can be limiting for certain tasks.

The CLIP [10] architecture is a state-of-the-art deep learning model that seeks to mitigate the limitations of pre-trained models concerning their real-world performance. It eliminates the requirement of task-specific training datasets and can effortlessly adapt to various tasks without arduous retraining the model. This is made possible by leveraging widely available large image-text datasets in its training process and attempting to match the visual perception of images with natural language. Despite its remarkable performance when compared to conventional ImageNet pre-trained models, CLIP still falls short in handling abstract or instance-specific tasks.

ALIGN [15] constitutes a similar architecture to CLIP, employing contrastive learning to match text-image pairs. Despite the architectural differences, ALIGN adopts a novel approach to data acquisition. Rather than subjecting the data to cleaning or filtering processes that could result in the presence of anomalous samples, ALIGN illustrates that the sheer magnitude of its data compensates for any potential noise, thereby incorporating a greater number of samples that would have otherwise been disregarded. LIT [16] augments CLIP and ALIGN by advancing a modular training paradigm that includes two distinct models - one for image embeddings, and the other for textual embeddings. The authors also introduce a novel technique referred to as contrastive tuning. This approach leverages contrastive pre-training with a pre-existing image model to detach the image model's learning process from image-text alignment, thereby improving the image features.

To further improve the generalization capabilities of the deep learning models the authors in [17] introduced Florence. They achieved this by incorporating task adaptations and the UniCL loss that integrates a single learning objective, enabling the seamless collaboration of two distinct data types (image-text). The task adaptations include mapping scenes to objects, images to videos and images to natural language. The data acquisition method is similar to ALIGN's. Florence significantly outperformed all the previous image-text foundation models in various vision transfer learning tasks.

The unification of image and text information presented a major challenge for foundational models in the image-text domain. ALBEF [18] attempted to address this issue by integrating the image and text spaces through cross-modal attention. The authors utilized the Image-Text Contrastive (ITC) loss to enhance the learning capability of the multimodal encoder while concurrently improving the hard negative mining selection process. Additionally, momentum distillation was employed to facilitate the model's learning from a larger dataset.

The Contrastive Captioners (CoCa) [19] represent a significant step forward in the image-text integration, addressing several limitations inherent to ALBEF regarding computational efficiency and performance. This model can be effectively trained from scratch, without any pre-trained weights, by employing a combination of contrastive and generative loss. The cross-attention mechanism is exclusively employed in the multimodal decoder layers to integrate the outputs from the image encoder. The captioning loss further enhances the integration by serving as a high-level, fine-grained descriptor. CoCa exhibits state-of-the-art results in various zero-shot tasks bridging the gap with supervised learning models.

B. METRIC LEARNING

Metric learning is a field of study that aims to develop discriminative descriptors, which are useful for classifying instances within datasets that have small inter-class variance. These datasets consist of objects that belong to the same category but are not the same instance (for example, different species of birds, cars, or types of food). There are various techniques for training models to perform tasks such as image retrieval and fine-grained classification. Many of these approaches focus on the loss function used during training, which can be divided into contrastive losses and classification losses.

Triplet loss [20] and N-pair loss [21] (InfoNCE [22], NT-Xent [23]) represent early endeavors in the field of contrastive loss functions. The former aims to reduce the proximity between positive pairs while concurrently maximizing the separation between negative pairs by utilizing triplets comprised of one positive and one negative example. The latter expands upon triplet loss by comparing a single positive example to multiple negatives, effectively generalizing the triplet loss concept. Both of these losses are sensitive to the selection process of positive and negative pairs, with the most favorable results being achieved through the use of hard-negative or semi-hard negative mining. Subsequently, a more advanced learning paradigm emerged, SupCon loss [24], which incorporated multiple positive and negative examples, thereby eliminating the need for fine-tuning the negative mining process. Another similar loss function is the multi-similarity loss [25], which uses the LogSumExp operation on

all pairs but gives special emphasis to the relative similarities between each embedding and its neighbors.

Classification losses include various modifications to the classic softmax loss. Normalized softmax [26] normalizes the input vectors to have unit magnitude, while Proxy-NCA [27] implements softmax on the Euclidean distances. Additionally, the angular-losses family, consisting of Arcface [12], CosFace [28], SphereFace [29], is a prevalent approach. ArcFace and CosFace compute the margin based on cosine similarity between the output embeddings and class weights, albeit they differ in their margin computation method. A variant of ArcFace, known as Sub-center Arcface [13], has been proposed to make the loss robust to label noise. Sphereface, on the other hand, transforms the features into a hypersphere space by computing the angle between the output vector and class weights.

III. THE PROPOSED APPROACH

A. MODEL ARCHITECTURE

Our approach utilized the CLIP architecture, which was initially proposed for the image-to-text task. Two different datasets were utilized for the training process in two public implementations - OpenAI's CLIP [10] and OpenCLIP [30]. The former involved pre-training the model on Imagenet22K, with ViT-L being the best-performing model. The latter involved training on the LAION-2B dataset [11], with ViT-H being the best-performing model. We only utilized the image encoder section of the original topology for both implementations. Our results indicated that the OpenCLIP model was significantly better than OpenAI, primarily because ViT-H had been exposed to more data and had greater learning capacity than ViT-L. Therefore, we proceeded with ViT-H CLIP.

The Vision Transformer (ViT) [31] is a new approach that uses the encoder part of the NLP Transformer [32] to process images. Images are split into fixed-sized patches and fed into the model with a learnable positional embedding vector assigned to every patch. The model uses self-attention with three components (Query, Key, and Value) to highlight important patches and a position-wise feed-forward neural network. The ViT model uses constant latent vector size in all its layers and incorporates skip connections and layer normalization. Multi-headed attention improves performance by allowing the model to focus on different positions and representation subspaces. The ViT-H14 model has 32 layers, a hidden dimension of 1280, an MLP size of 5120, 16 heads, and 632 million parameters.

We also modified the head by taking the output of the projection layer and feeding it to a BatchNormalization-Dropout-FullyConnected (FC) block. The dropout rate is set to 0.2, and the FC downsizes the 1024-dimensional embedding to 256. The 256 vector passes through the ArcFace layer to obtain the classification logits. Figure 2 showcases an overview of our proposed architecture.

To facilitate the instance retrieval task, we employed the Sub-Center ArcFace methodology. This approach has been suggested as a viable alternative to traditional facial recognition losses, as it demonstrates improved training convergence and performance by utilizing margin losses. The ArcFace methodology relies on the normalization of the dot product between the features extracted by a deep convolutional neural network (DCNN) and the last fully connected layer, resulting in a cosine distance calculation that allows for the computation of the angle between the feature and the centroid for each unique instance. An additional angular margin is then appended to each computed angle, and the resulting angles are converted to cosines before undergoing cross-entropy calculations. This variant of ArcFace enhances robustness when presented with noisy training data by relaxing the constraint that each class has a single center and introducing multiple sub-centers per class. Models trained using this approach can showcase improved representation capability and generate high-quality embeddings.

Despite the relatively even distribution of classes in our training dataset, some classes exhibit a disparity in representation, with some having fewer instances (approximately 5 images per class) in comparison to others (more than 40 images per class). To mitigate this issue, we utilized dynamic margins. Essentially, these margins are allocated dynamically for each class based on its number of instances. We assign larger margins to underrepresented classes to enhance the training procedure to facilitate their differentiation. The upper bound for the margins was established as 0.45, while the lower bound was set at 0.05.

B. DATASETS

We conducted a comprehensive examination of datasets closely aligned with the distribution of the test. To attain our optimal score, we selectively utilized subsets from the following datasets: Google LandmarksV2, Products10k, Food-101, iMaterialist, Fashion200k, DeepFashion, RP2K, Stanford Cars, Stanford Online Products, MET Artwork dataset, and Storefront-145. Despite the absence of instance labels in iMaterialist, we manually generated approximately 400 additional labeled furniture images with the assistance of the pre-trained CLIP. In an attempt to approximate the category percentages on the test set, we prioritized the most significant categories. During our preliminary experiments, we had roughly 200k images in the training set and observed that the dataset size and public leaderboard score displayed a corresponding trend. Hence, we amassed a dataset consisting of approximately 655k images from the aforementioned categories, randomly selecting classes with a minimum of three samples.

For the evaluation process, we selected a diverse array of instance recognition and retrieval datasets across a broad range of domains to demonstrate the universality of our fine-tuned model. To this end, we employed datasets from both domains familiar to the model through the training procedure and those unknown to it, including Paris6k, Oxford5k, INSTRE, In-Shop, Consumer-to-Shop, Cub200, Stanford Cars, UKBench, Inria Holidays, DukeMTMC-reID.

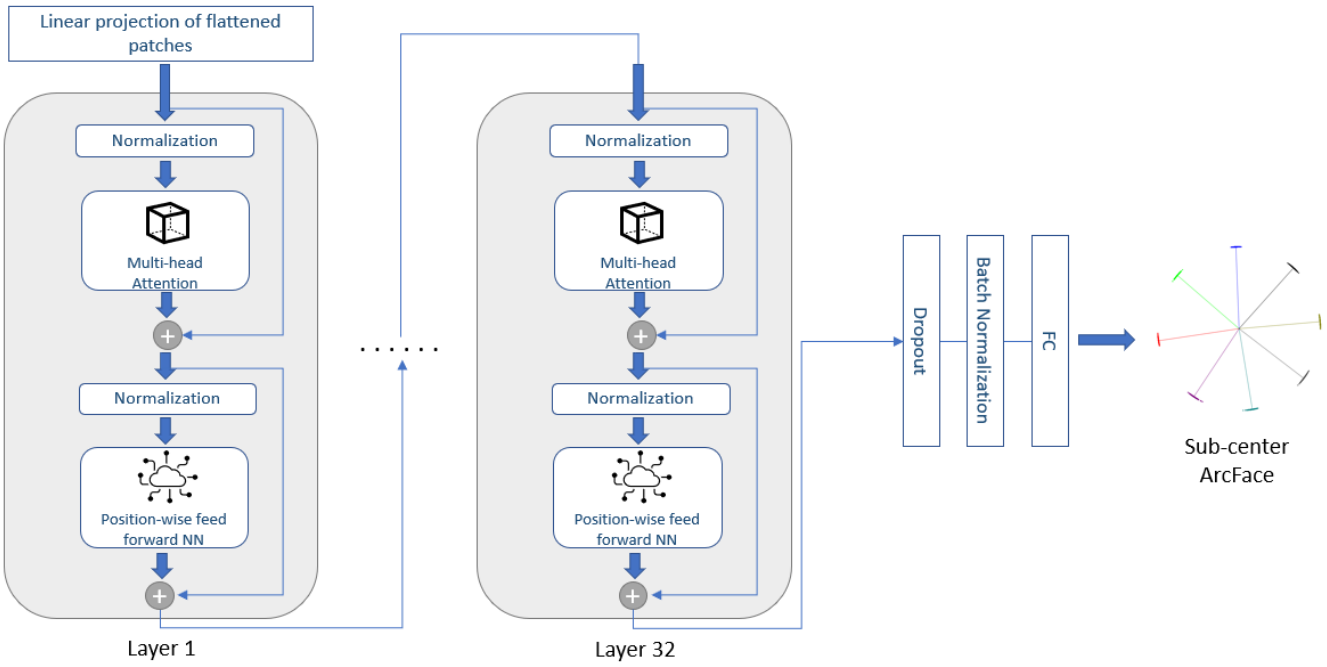

FIGURE 1. System architecture.

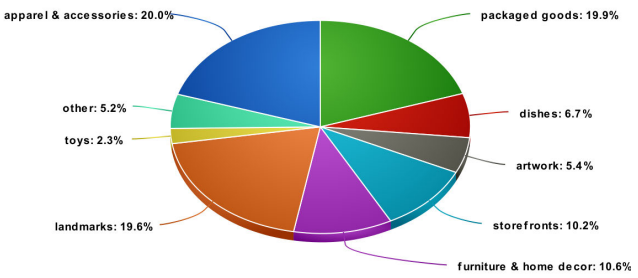

FIGURE 2. Test distribution of object types in the dataset of Kaggle competition.

- UKBench [33] comprises 10,200 images, divided into 2250 categories. Each category contains four pictures of the same object, captured from different angles and lighting conditions. To evaluate retrieval accuracy, the top-4 candidate score (NS score) is utilized for this dataset.
- INRIA Holidays [34] contains 1491 images captured by cellphones, depicting various holiday scenes and objects. The number of images per group ranges from 2 to 13. Unlike the UKBench database, the INRIA Holidays dataset offers numerous query images. The ground truth of the dataset includes images that match the query image in visual definition but does not specify whether they depict the same object or scene.
- The Paris6K Dataset [35] contains 6412 photographs that showcase distinct landmarks in Paris. Out of these, 55 images are dedicated to buildings and monuments that were requested to be included. Compared to the Oxford5K dataset, Paris6K exhibits a greater diversity of landmarks.
- The Oxford5K dataset [36] is comprised of 5062 images from Flickr that depict buildings. The dataset has been manually annotated to create a comprehensive ground truth for 11 unique landmarks, each of which is represented by five potential queries. The index comprises a total of 55 requests. Notably, different perspectives of the same building are labeled with the same name in this dataset, making it a challenging task for image retrieval.
- The INSTRE [37] dataset is utilized to evaluate various computer vision algorithms, including feature matching, invariant features, instance-level object retrieval, detection, and recognition. The dataset is divided into three separate subsets - INSTRE-S1 (single object case 1), INSTRE-S2 (single object case 2), and INSTRE-M (multiple object case). INSTRE-S1 consists of 11,011 images, and INSTRE-S2 consists of 12,059 images, with each subset having 100 distinct object classes. INSTRE-M comprises 5,473 images, grouped into 50 two-tuple classes, each featuring two distinct objects. The dataset categorizes the objects into architectures, daily stereoscopic objects, and planar objects.
- Stanford Cars [38]: The Cars repository encompasses 16,185 visual depictions, classified into 196 distinct categories of automobiles. The corpus is partitioned into two segments, with 8,144 images allocated for training and 8,041 designated for testing purposes, each category being apportioned equitably.

- The In-Shop Clothes Retrieval Benchmark constitutes a substantial portion of the DeepFashion [39] database. Substantial variations in pose and scale characterize this subset and exhibit a rich diversity of clothing items with a considerable quantity of annotated images. Specifically, it comprises 7,982 articles of clothing, 52,712 in-shop clothing images, and approximately 200,000 cross-pose/scale pairs. Additionally, each image is comprehensively annotated with bounding box information, garment classification, and pose classification.
- The Consumer-to-Shop Clothes Retrieval Benchmark is also a substantial component of the DeepFashion database that evaluates the competency of a retrieval system to match consumer photos of apparel to the corresponding shop photo. This benchmark is distinguished by the presence of cross-domain correspondences and diverse variations encountered in real-world scenarios. It encompasses a substantial number of clothing items, totaling 33,881, and an abundant quantity of consumer and shop-clothing images, numbering 239,557, along with 195,540 cross-domain pairs. Annotations are similar to In-shop.
- The CUB-200 dataset [40], also known as the Caltech-UCSD Birds-200, is a comprehensive image dataset of birds that contains 200 species of birds and more than 11,000 images. The dataset was created to train computer vision algorithms and evaluate their performance in recognizing and classifying birds. The images in the dataset were taken from various angles, under different lighting conditions, and in various environments.
- Google landmarks v2 (GLDv2) dataset [41] is a large-scale instance recognition/image retrieval dataset that features more than 5M images with 200k different classes. GLDv2 dataset poses significant difficulties for researchers due to the disparity in the distribution of samples across its different classes and the significant variance within each class.
- The RP2K dataset [42] is a recent and extensive collection of instance recognition data for the retail industry, comprising over 500,000 images and featuring 2000 distinct categories. The images were acquired through manual means within physical retail stores, under conditions that accurately reflect reality.
- The Food Recognition Dataset [43] contains 101,000 images of 101 food categories, created to evaluate the performance of a Random Forest-based technique for extracting discriminative parts of images. The images were pre-processed only to consider patches aligned with image superpixels, called components, for improved efficiency in mining and classification.
- Another large-scale dataset in the retail sector is the Products-10k [44] which is human-labeld and features 10k products. The products belong to diverse categories: fashion, furniture, packaged goods, and food. Labels are provided at both instance and category levels.
- The Stanford online products dataset [45] comprises around 120,000 images and over 22,000 classes. The images have been sourced from eBay and primarily consist of objects related to home appliances and furniture.
- Fashion200k [46] is another dataset created by crawling various online shopping websites to collect fashion-related images and their descriptions. The dataset includes five clothing categories: dress, top, pants, skirt, and jacket. The purpose of this dataset is to develop a visual-semantic embedding.
- Storefront-145[2] is a collection of 4,545 retail storefront images generated by a Kaggle user. The images are divided into 145 storefront brands.
- DukeMTMC-reID [47] is a dataset created by assembling high-resolution videos captured by eight cameras. The dataset includes footage of pedestrians, which has been manually cropped. The dataset comprises 16,522 images and 702 pedestrian instances.
- Met [48] is an art collection dataset sourced from the Metropolitan Museum of Art. It comprises around 400,000 images and 224,000 instances. Like the Google landmarks dataset, Met also exhibits a disparity in the distribution of samples across its various classes, with more than half of the classes containing only one sample.
- iMaterialist[3] is another Kaggle dataset that was created as part of the Fifth Workshop on Fine-Grained Visual Categorization(FGVC5) workshop for fine-grained visual categorization. It contains over 210k images from 128 furniture and home decor classes.

C. EXPERIMENTS

The OpenCLIP model offers top-notch performance on zero-shot tasks, enabling it to provide high-quality image descriptors without additional training. However, training the model without causing catastrophic forgetting was challenging as significant changes to the learned weights can cause the loss of previously acquired knowledge. During our initial training attempts, we kept the CLIP backbone frozen and trained the head for five epochs with a learning rate of 0.0001. Subsequently, we unfroze the ViT-H until resblock 15 and trained for an additional epoch with the learning rate reduced by a factor of 10. As we incorporated more data into our pipeline, we realized we could further train the backbone. To address this, we adopted a scheme that involves applying different learning rates to the backbone and the head.

$$LearningRate = \begin{cases} 1e^{-7}, & for\ model.backbone \\ 1e^{-4}, & for\ model.head \end{cases}$$

and trained the model for four epochs. We utilized a batch size of 32 and employed the Adam optimizer. We limited the model's unfreezing to resblock 31 to resblock 15 for three primary reasons. Firstly, we did not observe any improvement

[2]https://www.kaggle.com/datasets/kerrit/storefront-146
[3]https://www.kaggle.com/c/imaterialist-challenge-furniture-2018

by unfreezing additional blocks. Secondly, we considered the computational complexity and training time associated with unfreezing more blocks. Finally, we noted that the initial layers typically contain low-level generic features, and we did not want to modify the learned features from a larger dataset.

We used the following augmentations during training inspired by past Kaggle instance-level competitions: Horizontal flip, image compression, shift, scale, rotate, cutout, random brightness, contrast, and RGB-shift.

IV. EVALUATION RESULTS

To exhibit the refined generative capacity of CLIP, we have incorporated a diverse range of instance recognition and retrieval tasks and have contrasted our results with those obtained from the pre-trained CLIP and with the current top-performing models for each benchmark. The assessment datasets can be divided into two categories: relevant datasets to the training domain - that is, domains that have been encountered during the training procedure; and datasets that encompass uncharted domains, where the model has not been directly fine-tuned to separate instances.

The first category comprises of Paris, Oxford, consumer-to-shop, Stanford cars, datasets similar to what the model has been trained on, such as Google landmarks, Stanford training set, and In-Shop dataset, which includes apparel. The rest of the datasets belong to the second category as they either contain generic instances such as UKBench, Inria, INSTRE or unknown instances such as CUB-200 and DukeMTMC-reID,

The model was refined during Kaggle's competition to yield optimal results on Kaggle's private dataset. In this manuscript, the model has been employed out-of-the-box, without any additional tweaks to enhance its performance on the utilized benchmarks. This constitutes another notable advantage of our approach, which deviates from the conventional practice of fine-tuning a method by getting feedback on the evaluation benchmarks, as is often observed in published results.

For inference, we resized the images to 224×224 to match the expected input of the model and then normalized them. Feature extraction was carried out by utilizing the projection layer located after the final resblock of ViT-H14, which had an embedding size of 1024. In our specific scenario, we also incorporated the arc face-refined descriptor with a dimensionality of 256 and the PCA instance, which we had previously used in Kaggle's competition, to reduce the embedding size from 256 to 64. The PCA was trained on a public dataset[4] consisting of 130k images from all categories of the test set, which was shared on Kaggle.

Tables 1 and 2 showcase the performance of our approach, as well as the original clip and the state-of-the-art (SOTA) result for each individual benchmark. It's important to note that we cherry-picked the best-performing algorithm for each distinct domain based on reported results to compare with our approach. This means that we compared our method to multiple algorithms that have set the SOTA result in each domain, emphasizing the importance of our performance evaluation. The SOTA row in the results table represents the highest reported result for each domain.

To elaborate, in studies [50] and [49], researchers focused on effectively merging local and global features within CNNs and Transformer architectures to bolster their performance. In [47], a novel approach is introduced to enhance Vision Transformer embeddings by applying hyperbolic geometry to these structures. In [53] and [55], the authors developed methods to capitalize on the nearest neighbor graph, which retains discriminative information. The first technique utilizes a convolutional graph network to seamlessly integrate graph data into the embedding. In contrast, the second method presents a manifold-ranking procedure accompanied by an innovative graph architecture. Consequently, both approaches effectively employ the nearest neighbor graph to elevate the overall performance of their respective models. Lastly, the research in [51] investigates the use of class centroids for retrieval by implementing a novel centroid triplet loss function, while in [54], the authors trained a model to determine the optimal method to complement the feature output of a baseline model and subsequently fuse the acquired information.

Comparing the CLIP model and our finetuned variant on Table 1 and 2, we observe a significant enhancement in results across all benchmarks consequent to finetuning. Notably, the finetuned model substantially increases performance for datasets within similar domains and improves upon generic, unknown domains. In view of the likelihood of catastrophic forgetting in such circumstances, which can lead to a distributional shift, our findings attest to the effectiveness of the proposed training procedure. The results prove that the model leverages its pre-existing learned features and expands its knowledge to new domains without any forgetting occurring.

The second observation that can be extrapolated from the figures pertains to the performance of our method relative to the SOTA. Even though the SOTA outcomes presented herein are derived from explicit training on pertinent domains for each benchmark while also applying other advanced techniques(feature fusion, re-ranking, etc.) that improve performance substantially, our model exhibits comparable and even superior performance in certain cases, such as Consumer-to-shop and Stanford Cars. Our approach demonstrates notable competency even on unfamiliar domains, such as CUB, Duke, and Instre. It is worth noting that our training set did not contain a single image from these datasets or similar classes, highlighting the generalization ability of our method. It stands to reason that the difference between domain-specific and domain-agnostic models is relatively small across a broad range of domains, and with the incorporation of a larger dataset, the disparity would likely be negligible.

It is important to highlight that the comparison tables presented in our paper do not directly aim to compare our method with SOTA benchmarks. Instead, we demonstrate that

[4]https://www.kaggle.com/datasets/rhtsingh/google-universal-image-embeddings-128 × 128

TABLE 1. Performance evaluation of fine-tuned CLIP on general or new domains compared to the original CLIP and SOTA methods. The first column also includes the corresponding embedding size. Specifically, 1024 refers to the size of the last projection layer, 256 refers to the arcface refined layer, and 64 refers to the PCA.

Model	Paris	Oxford	Consumer-to-shop				Stanford Cars			
-	mAP	mAP	R@1	R@10	R@20	R@50	R@1	R@2	R@4	R@8
$CLIP^{1024}$	80.39	74.81	19.19	42.67	49.70	59.81	91.47	95.98	98.21	**98.91**
$CLIP^{1024}$ (ours)	88.12	87.69	49.19	80.28	85.06	90.23	94.29	97.08	**98.21**	98.89
$CLIP^{256}$ (ours)	88.98	89.66	**51.57**	**82.81**	**87.42**	**92.01**	**94.75**	**97.19**	98.13	98.71
$CLIP^{64}$ (ours)	87.44	85.44	47.22	77.96	83.07	88.37	94.42	96.82	98.02	98.69
SOTA	**97.60** [49]	**95.50** [50]	29.40 [51]	61.30 [51]	68.90 [51]	77.40 [51]	89.2 [52]	94.1 [52]	96.7 [52]	98.1 [52]

TABLE 2. Perfomance evaluation of fine-tuned CLIP on general or new domains compared to the original CLIP and SOTA methods. The first column also includes the corresponding embedding size. Specifically, 1024 refers to the size of the last projection layer, 256 refers to the arcface refined layer, and 64 refers to the PCA.

Model	INSTRE	Inria	UKBench	Duke			CUB			
-	mAP	mAP	N-4	R@1	R@5	R@10	R@1	R@2	R@4	R@8
$CLIP^{1024}$	73.17	91.74	3.89	60.82	76.80	81.51	82.11	89.30	93.92	96.62
$CLIP^{1024}$ (ours)	78.86	93.09	3.92	65.75	79.98	84.69	83.37	89.95	94.32	96.56
$CLIP^{256}$ (ours)	82.01	93.05	**3.93**	65.22	79.13	83.08	84.13	90.39	94.30	96.59
$CLIP^{64}$ (ours)	69.76	89.48	3.89	53.59	69.48	75.09	81.32	88.54	93.24	95.95
SOTA	**89.20** [53]	**96.10** [54]	**3.93** [55]	**95.60** [51]	**96.20** [51]	**97.90** [51]	**85.60** [52]	**91.40** [52]	**94.80** [52]	**96.70** [52]

a single model with a single fine-tuning and a single descriptor can deliver impressive results across diverse domains, approaching or surpassing SOTA techniques' performance. It is worth noting that many SOTA methods have been tailored to achieve optimal performance in specific benchmarks, which could potentially lead to overfitting for those particular tasks. Our approach, on the other hand, emphasizes the model's versatility and adaptability to handle various domains, providing a practical solution for transfer learning across multiple domains.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method that builds on the pre-existing CLIP architecture on LAION-2B dataset to create a universal image encoder that can be fine-tuned for tasks such as image retrieval and recognition. We demonstrate that our method outperforms the original CLIP in all benchmarks, including those in familiar and unfamiliar domains, and can even match or surpass the performance of state-of-the-art algorithms trained explicitly for a single domain. The results suggest that our approach offers a practical solution for transfer learning that addresses the challenges of dealing with multiple domains. The implications of this research are significant for real-life applications where multiple domains are common.

One significant limitation that may arise in this setup is the high number of distinct instance classes involved when multiple domains are considered. This problem is twofold: first, it can become computationally burdensome as the fully-connected layer grows in size, and second, instance separability can become increasingly difficult as the number of domains and instance classes increases. A larger dataset with even more domains would need to be used to understand this limitation better. Unfortunately, such a dataset does not currently exist.

REFERENCES

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[2] S. Gkelios, A. Sophokleous, S. Plakias, Y. Boutalis, and S. A. Chatzichristofis, "Deep convolutional features for image retrieval," *Exp. Syst. Appl.*, vol. 177, Sep. 2021, Art. no. 114940.

[3] S. Gkelios, Y. Boutalis, and S. A. Chatzichristofis, "Investigating the vision transformer model for image retrieval tasks," in *Proc. 17th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, 2021, pp. 367–373.

[4] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.

[5] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (BiT): General visual representation learning," in *Computer Vision—ECCV* (Lecture Notes in Computer Science). Glasgow, U.K.: Springer, 2020, pp. 491–507.

[6] Z. Xiao, H. Zhang, H. Tong, and X. Xu, "An efficient temporal network with dual self-distillation for electroencephalography signal classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2022, pp. 1759–1762.

[7] H. Xing, Z. Xiao, D. Zhan, S. Luo, P. Dai, and K. Li, "SelfMatch: Robust semisupervised time-series classification with self-distillation," *Int. J. Intell. Syst.*, vol. 37, no. 11, pp. 8583–8610, Nov. 2022.

[8] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, and B. Zhao, "A federated learning system with enhanced feature extraction for human activity recognition," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107338.

[9] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 7959–7971.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[11] C. Schuhmann, R. Beaumont, C. W. Gordon, R. Wightman, T. Coombes, A. Katta, C. Mullis, P. Schramowski, S. R. Kundurthy, and K. Crowson, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Proc. 26th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2022, pp. 1–50.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[13] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, "Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces," in *Computer Vision—ECCV* (Lecture Notes in Computer Science). Glasgow, U.K.: Springer, 2020, pp. 741–757.

[14] Q. Ha, B. Liu, F. Liu, and P. Liao, "Google landmark recognition 2020 competition third place solution," 2020, *arXiv:2010.05350*.

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[16] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "LiT: Zero-shot transfer with locked-image text tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 18123–18133.

[17] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, and C. Li, "Florence: A new foundation model for computer vision," 2021, *arXiv:2111.11432*.

[18] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 9694–9705.

[19] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," 2022, *arXiv:2205.01917*.

[20] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, no. 2, pp. 207–244, 2009.

[21] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[22] A. Van Den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[25] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5022–5030.

[26] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L_2 hypersphere embedding for face verification," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1041–1049.

[27] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 360–368.

[28] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[29] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 212–220.

[30] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," 2022, *arXiv:2212.07143*.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[33] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search," in *Proc. 19th ACM Int. Conf. Multimedia*, New York, NY, USA, Nov. 2011, pp. 1437–1440, doi: 10.1145/2072298.2072034.

[34] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2008, pp. 304–317.

[35] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[36] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[37] S. Wang and S. Jiang, "INSTRE: A new benchmark for instance-level object retrieval and recognition," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 11, no. 3, pp. 1–21, 2015.

[38] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. 4th Int. IEEE Workshop 3D Represent. Recognit.*, Mar. 2013, pp. 554–561.

[39] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[40] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD birds 200," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, 2010. [Online]. Available: https://resolver.caltech.edu/CaltechAUTHORS:20111026-155425465

[41] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2—A large-scale benchmark for instance-level recognition and retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2575–2584.

[42] J. Peng, C. Xiao, and Y. Li, "RP2K: A large-scale retail product dataset for fine-grained image classification," 2020, *arXiv:2006.12634*.

[43] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101—Mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 446–461.

[44] Y. Bai, Y. Chen, W. Yu, L. Wang, and W. Zhang, "Products-10K: A large-scale product recognition dataset," 2020, *arXiv:2008.10545*.

[45] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[46] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic spatially-aware fashion concept discovery," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1463–1471.

[47] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV Workshops*, 2016, pp. 17–35.

[48] N.-A. Ypsilantis, N. Garcia, G. Han, S. Ibrahimi, N. V. Noord, and G. Tolias, "The met dataset: Instance-level recognition for artworks," in *Proc. 34th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=fnuAjFL7MXy

[49] C. H. Song, J. Yoon, S. Choi, and Y. Avrithis, "Boosting vision transformers for image retrieval," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 107–117.

[50] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang, "DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11772–11781.

[51] M. Wieczorek, B. Rychalska, and J. Dabrowski, "On the unreasonable effectiveness of centroids in image retrieval," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2021, pp. 212–223.

[52] A. Ermolov, L. Mirvakhabova, V. Khrulkov, N. Sebe, and I. Oseledets, "Hyperbolic vision transformers: Combining improvements in metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7409–7419.

[53] C. Liu, G. Yu, M. Volkovs, C. Chang, H. Rai, J. Ma, and S. K. Gorti, "Guided similarity separation for image retrieval," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[54] C. Malone, S. Hausler, T. Fischer, and M. Milford, "Boosting performance of a baseline visual place recognition technique by predicting the maximally complementary technique," 2022, *arXiv:2210.07509*.
[55] G. Lao, S. Liu, C. Tan, Y. Wang, G. Li, L. Xu, L. Feng, and F. Wang, "Three degree binary graph and shortest edge clustering for re-ranking in multi-feature image retrieval," *J. Vis. Commun. Image Represent.*, vol. 80, Oct. 2021, Art. no. 103282.

SOCRATIS GKELIOS received the Diploma degree from the Department of Electrical and Computer Engineering, Democritus University of Thrace, in 2019, where he is currently pursuing the Ph.D. degree in computer science. He is a Research Assistant with the Centre for Research and Technology Hellas (CE.R.T.H.), Information Technologies Institute (ITI). His research interests include computer vision and deep learning.

ANESTIS KASTELLOS received the bachelor's degree from the Department of Mathematics, University of Ioannina, in 2021. He is currently pursuing the M.Sc. degree in artificial intelligence with the Department of Computer Science, European University of Cyprus. He is a Research Assistant with the Centre for Research and Technology Hellas (CE.R.T.H.), Information Technologies Institute (ITI) and a Research Assistant at Intelligent Systems Laboratory, Neapolis Unversity Pafos, Cyprus. His research interests include computer vision, robotics, and deep learning.

YIANNIS S. BOUTALIS (Senior Member, IEEE) received the Diploma degree in electrical engineering from the Democritus University of Thrace (DUTH), Greece, in 1983, and the Ph.D. degree in electrical and computer engineering from the Computer Science Division, National Technical University of Athens, Greece, in 1988. Since 1996, he has been a Faculty Member with the Department of Electrical and Computer Engineering, DUTH, where he was a Lecturer, an Assistant Professor, and an Associate Professor. He is currently a Professor, the Director of the Automatic Control Systems and Robotics Laboratory, and the Head of the Department of Electrical and Computer Engineering, DUTH. In the past, he served as an Assistant Visiting Professor with the University of Thessaly (PD 407), Greece, and a Visiting Professor (under contract) with the Air Defense Academy of General Staff of Air Forces, Greece. His current research interests include developing computational intelligence techniques with applications in control, renewable power generation, electric machines, robotics, pattern recognition, and signal and image processing problems.

SAVVAS A. CHATZICHRISTOFIS (Senior Member, IEEE) received the Diploma and Ph.D. degrees (Hons.) from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece. Since 2017, he has been a Faculty Member with the Department of Computer Science, Neapolis University Pafos, where he is currently a Professor, the Vice-Rector of research and innovation, and the Director of the Intelligent Systems Laboratory (ISLab). His research interests include the intersection of artificial intelligence, computer vision, robotics visual feature extraction, image analysis, matching, indexing and retrieval, SLAM, and educational robotics. He has more than 15 years of solid experience and reporting more than 100 publications in these fields. He has received several distinctions, grants, scholarships, and awards for his research contribution.