Enhancing Remote Sensing Image Retrieval: A Hierarchical Approach Integrating Visual and Semantic Similarities

Wen Lu, Minh Nguyen*

***Abstract*—The heightened revisiting frequency and expanded observation capabilities of satellites lead to the daily generation of a substantial volume of remote sensing images. Retrieving relevant data accurately from this extensive archive holds significant importance. Deep learning Content-Based Image Retrieval (CBIR) utilizes a feature extraction network pre-trained on image classification tasks to derive image-level features. Subsequently, a similarity measure is applied on these features to identify the archive images most closely resembling the query image. While image-level labels facilitate CBIR in retrieving images from the same category as the query image, they do not empower CBIR to differentiate between implicit sub-categories. For instance, although CBIR can discern between broader categories like "residential" and "forest", it lacks the necessary semantic statistical information to distinguish more nuanced distinctions such as "high-density residential" from "medium-density residential". To enhance image retrieval for greater similarity, we propose a Hierarchical Image Retrieval (HIR) approach that combines visual similarity with semantic statistics. In the first stage, visually similar images are identified using CBIR, while the second stage refines the selection based on semantic similarity derived from land-cover classification. The experimental results indicate that HIR achieves 20% higher retrieval accuracy for "residential" sub-categories and over 1% increase in retrieval accuracy across all classes.**

***Index Terms*—image retrieval, land cover, remote sensing.**

I. Introduction

PRESENTLY, the world has established a high-resolution Earth observation system utilizing aerospace remote sensing technology, significantly enhancing the efficiency and capacity for acquiring remote sensing data. Numerous remote sensing satellites in orbit have the capability to gather extensive image data swiftly, supporting various applications like resource surveying and mapping [1]. However, this surge in data acquisition also presents challenges in effectively managing and retrieving this vast volume of information. Image retrieval involves gathering pertinent images from the archive, priming the data for subsequent tasks such as image matching, registration, and fusion.

Text-Based Image Retrieval (TBIR) necessitates the manual assignment of keywords to each image, thereby establishing image metadata, and subsequently organizing it within a database management system. This system conducts searches within the database using user-entered keywords to retrieve relevant images. Nevertheless, remote sensing images encompass intricate textures and complex geometric structures, posing challenges in accurately encapsulating their content through keyword tags. Moreover, the exponential surge in data volume renders the increasingly time-consuming and costly manual annotation of images impractical. Instead of relying on metadata or textual descriptions, Content-Based Image Retrieval (CBIR) analyzes the visual features of images themselves to perform searches, thereby overcoming the limitations of TBIR [2]. Deep learning CBIR utilizes a feature extraction network pre-trained on image classification tasks to derive image-level features. Subsequently, a similarity measure is applied on these features to identify the archive images most closely resembling the query image. During the pre-training phase of the feature extraction network, every image in the image classification dataset is annotated by single or multiple broader category image-level labels. These labels serve to represent the most significant content depicted within the images. While broader category image-level labels facilitate CBIR in retrieving images of the same category as the query image, they do not empower CBIR to differentiate between implicit sub-categories. For instance, CBIR can distinguish between broader categories like "residential" and "forest", but it falls short in discerning subtleties such as distinguishing between "high-density residential" and "medium-density residential". The first row of Figure 5 demonstrates that only 3 out of the top 10 images retrieved by CBIR belong to "medium-density residential" as the query image. The distribution, proportion, and function of semantic objects determine the various land-use types [3], therefore, statistical information becomes imperative in subdividing explicit broader categories into nuanced implicit sub-categories in an unsupervised manner.

In this study, we propose using land cover classification to incorporate statistical data into image retrieval processes. Land cover classification is a multi-class segmentation task where each pixel is categorized into natural or man-made elements on the Earth's surface, such as water, soil, natural vegetation, crops, and human infrastructure. Analyzing a remote sensing image through land cover classification enables the calculation of the proportion of pixels in each specific land cover category. By assessing the ratio of pixels categorized as "building", algorithms can determine if the remote sensing image corresponds to implicit sub-categories like "high-density residential" or "medium-density residential". While land cover classification provides valuable statistical information, it solely captures se-

Wen Lu and Minh Nguyen are with School of Engineering, Computer & Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand (e-mail: wen.lu@autuni.ac.nz, minh.nguyen@aut.ac.nz).

*, Minh Nguyen is the corresponding author.

mantic information, lacking visual content information essential for distinguishing architectural styles. For instance, both villas and high-rise apartment buildings fall under the same land cover category, "residential", making it impossible for Land Cover-Based Image Retrieval (LCBIR) to differentiate between them. Consequently, images retrieved via LCBIR may not visually resemble the query image. Additionally, land cover classification is a pixel-wise dense prediction task using a semantic segmentation network, demands extensive computational resources. Overall, both CBIR and LCBIR have limitations. Only by combining visual and semantic similarities can we enhance the retrieval of more closely related remote sensing images. Consequently, we propose a two-stage Hierarchical Image Retrieval (HIR) approach that supplements visual similarity with semantic statistics.

HIR introduces an additional stage, thereby introducing additional computational complexity. Hence, an efficient feature extraction network named TKANet was developed to reduce the processing time of Stage 1, thereby offsetting the additional time required for Stage 2. TKANet employs a repeatable module: it initially diffuses into three parallel branches for multi-scale feature extraction and subsequently converges by adding feature maps without expanding the channel count. This design facilitates faster inference compared to current lightweight networks while generating more distinctive features. In addition, to augment the retrieval accuracy of our hierarchical approach, we designed a Transformer model with multi-granularity tokens, which has demonstrated promising accuracy in land cover classification tasks.

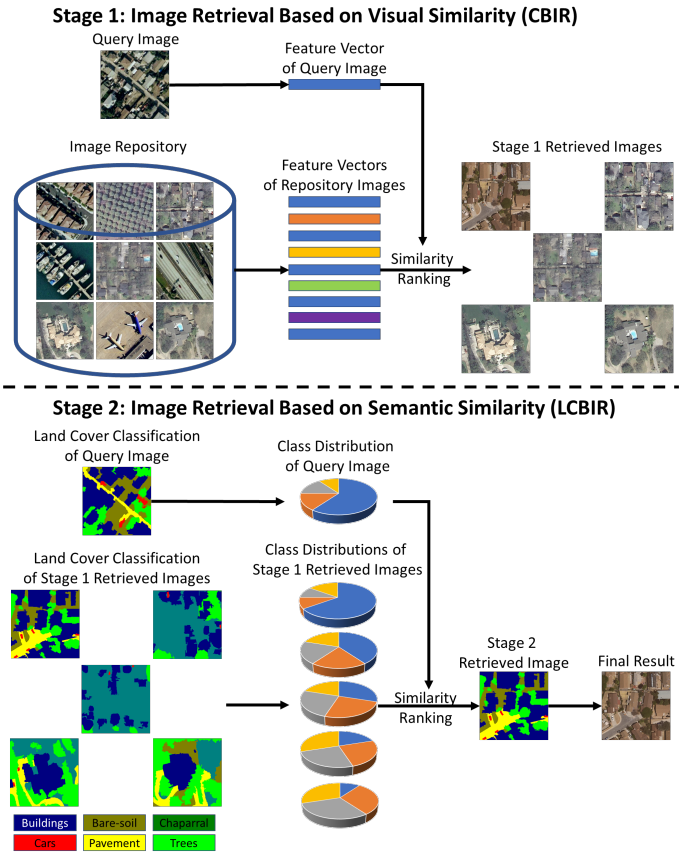

Fig. 1. The Flowchart of Two-Stage Hierarchical Image Retrieval Approach.

II. Methodology

A. Two-Stage Hierarchical Image Retrieval Approach

As depicted in Figure 1, our proposed HIR operates in two stages. Stage 1 involves CBIR, gathering visually similar images, followed by Stage 2, which employs LCBIR to refine the retrieval results based on semantic similarity.

The upper section of Figure 1 illustrates the first stage, where repository images are input into a feature extraction network, generating image-level feature vectors stored for reference. Upon initiating a retrieval request, the query image is processed through the feature extraction network to produce its own image-level feature vector. Utilizing a distance measure (e.g., Euclidean Distance, Manhattan Distance, or Cosine Distance), the system calculates the distances between the query image's feature vector and those of each repository image. These distances are then sorted in ascending order. The repository images associated with the top K smallest distances constitute the Stage 1 retrieval results, placed on a shortlist as candidates for Stage 2.

The lower section of Figure 1 illustrates the second stage, where a land cover classification network predicts the land cover class of each pixel within both the query image and the candidate images shortlisted in the previous stage, generating a distribution of pixel classes for each image. Subsequently, the Manhattan Distance metric is applied to determine the distances between the class distribution of the query image and those of the images within the shortlist. These distances are then ranked, and the images corresponding to the top k (where $k < K$) smallest distances constitute both Stage 2 and the final retrieval results.

Class weights can be applied in Manhattan Distance calculation to emphasize some land-cover classes. For instance, in scenarios where the query image portrays a residential area adjacent to a river or lake, augmenting the weight of the "water" land-cover class can elevate the ranking of images featuring water bodies. Weighted Manhattan Distance (WMD) between two land-cover class distributions $\boldsymbol{X} = (x_1, x_2, ..., x_n)$ and $\boldsymbol{Y} = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$ is defined as:

$$\mathbf{WMD}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{n} w_i |x_i - y_i| \tag{1}$$

where n is the number of land-cover classes, and $\boldsymbol{w} = (w_1, w_2, ..., w_n)$ denotes class weights.

Conventional CBIR methods harbor two notable shortcomings that can lead to inaccurate retrieval. One issue involves an over-sensitivity to a specific class while disregarding coexisting classes within the image. For instance, when the query image displays a river flanked by grass or trees, a CBIR method sensitive to plants might retrieve images of grasslands or forests. This arises from the feature extraction network's training on datasets annotated with single labels representing the most prominent content. In cases where images contain multiple land-cover categories or object classes,

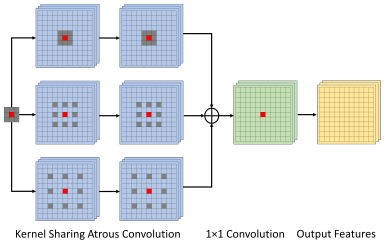

Fig. 2. TKA comprises two tiers of tri-branch 3×3 kernel-sharing depthwise atrous convolutions using dilation rates (1, 2, 3), followed by feature map fusion through addition, and channel fusion via pointwise convolution.

the network tends to prioritize features linked to only one category or class. The other shortcoming lies in the high intra-class variation and low inter-class differences prevalent in remote sensing imagery, often causing CBIR to erroneously gather images from categories different from that of the query image. Our hierarchical approach effectively addresses the two shortcomings by leveraging additional land-cover statistical information, enabling LCBIR in Stage 2 to eliminate mis-collected images. In the aforementioned example, images devoid of any water content are filtered out during Stage 2, enhancing the accuracy of image retrieval through this double-check procedure. Moreover, compared to the extensive archive of images, our approach involves inputting a minimal proportion of images (for example, $K = 10k$) into the semantic segmentation network. As a result, our HIR proves significantly more efficient than LCBIR while maintaining comparable computational complexity with CBIR.

Unlike multi-labeled CBIR, which handles semantic information imprecisely and suffers from the absence of ground truth images for each query image, thus lacking protocols for similarity measures and performance evaluations [4], HIR Stage 2 quantitatively analyzes semantic information and employs the same evaluation metrics as single-labeled CBIR.

B. Low-dimensional Feature Extraction Network for Stage 1

We employ the Tri-branch Kernel-sharing Atrous Convolution Module (TKA), illustrated in Figure 2, as the fundamental module to extract features from multi-scale ground objects.

The network architecture of TKANet is outlined in Table I. The 32-dimensional vectors produced by the fully-connected layer serve as image-level features for Stage 1 CBIR. In contrast to other lightweight networks, which yield feature vectors exceeding 1000 dimensions, TKANet significantly reduces storage requirements and computation costs to 1/32. This alleviates the challenges associated with large-scale image retrieval and frequent updates to image repositories.

C. Land Cover Classification Network for Stage 2

As shown in Figure 3, our land cover classification network, ESFormer, employs the EfficientNetV2 backbone as the encoder for extracting multi-scale features. The output features of EfficientNetV2 Stage 5 undergo nearest interpolation for upsampling and are then combined with Stage 3 output features. Subsequent to linear projections, $\mathbf{Q}'$, $\mathbf{K}'$, and $\mathbf{V}'$ are

TABLE I
THE NETWORK ARCHITECTURE OF TKANET.

Stage	Output Size	Operation	Output Channels
Input Image	256×256		3
Stage 1	128×128	ConvS2	32
Stage 2	64×64	ConvS2	64
Stage 3	32×32	ConvS2	128
	32×32	TKA module	128
Stage 4	16×16	ConvS2	192
	16×16	TKA module	192
Stage 5	8×8	ConvS2	256
	8×8	TKA module	256
Pooling	1×1	GAP	256
Features	1	Fully-connected	32

ConvS2: 3×3 convolution with stride 2, batch normalization, ReLU. GAP: Global Average Pooling.

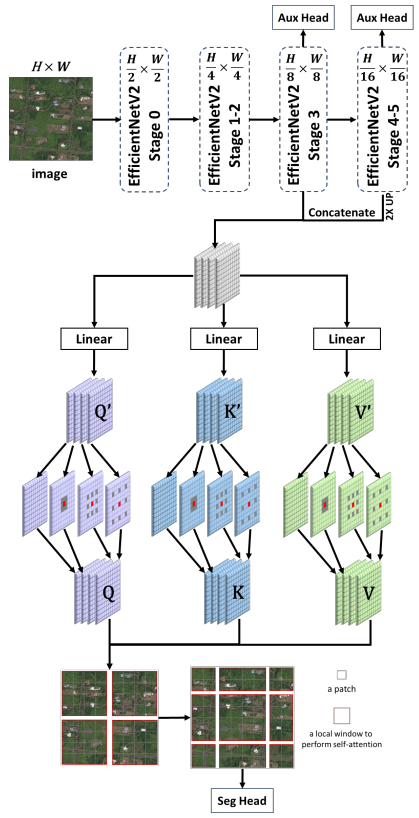

Fig. 3. ESFormer employs the EfficientNetV2 backbone as encoder, utilizes the Multi-Granularity Tokens and Swin Transformer blocks as decoder.

divided into four token segments. The initial segment remains unaltered to preserve its representation of token-to-token relationships in subsequent multi-head self-attention. Depthwise convolutions are subsequently applied to the following three segments, each with distinct dilation rates (0, 1, 2). This approach aims to amalgamate token segments related to 3×3, 5×5, and 7×7 patches, respectively, to capture patch-to-patch relationships in ensuing multi-head self-attention. These four segments, possessing varied receptive fields, are then concatenated to form the Multi-Granularity Tokens $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ before being passed to sequential Swin Transformer blocks. A 1×1 convolutional layer serves as the semantic segmentation

TABLE II
STAGE 1 CBIR RESULTS OF VARIOUS FEATURE EXTRACTION NETWORKS

Method	mAP@10	mAP@20	P@10	P@20	Speed
ShuffleNetV2	0.9030	0.8896	0.8804	0.8591	353
STDC1	0.9119	0.8972	0.8848	0.8617	574
REGNETX400	0.9096	0.8995	0.8876	0.8723	297
MobileNetV3	0.9045	0.8978	0.8901	0.8811	456
MobileNetV2	0.9282	0.9194	0.9063	0.8983	451
EfficientNetB0	0.9288	0.9193	0.9071	0.8950	289
ResNet18	0.9478	0.9360	0.9246	0.9105	490
TKANet	**0.9532**	**0.9502**	**0.9457**	**0.9434**	**1322**

Inference Speeds (FPS) were measured on NVIDIA RTX3050 Mobile GPU.

head, accepting the decoder's output features. The resulting logits undergo an 8X upsampling through bilinear interpolation before being fed into the primary loss function. To bolster the encoder's feature extraction capability, two auxiliary heads are introduced atop Stage 3 and Stage 5 output features during the training phase. During inference, these auxiliary heads are discarded, incurring no additional computational cost. The output logits of these auxiliary heads undergo 8X and 16X upsampling, respectively, through bilinear interpolation before being fed into two auxiliary loss functions.

III. EXPERIMENTS AND ANALYSIS

A. Experimental Setup

DLRSD dataset [5] has both image-level labels (21 categories, 100 images per category) and pixel-level labels (17 classes), therefore, it can be used for image classification and retrieval tasks as well as semantic segmentation task. Our study aimed to assess whether Stage 2 of our hierarchical approach could refine Stage 1 retrieval outcomes to images sharing the same residential density as the query image. For this purpose, we merged "Dense Residential", "Medium Residential", and "Sparse Residential" into a unified category termed "Residential", withholding the subcategory labels from the networks during the training process. As previous studies, the total 2100 images were divided into 1470 archive images, also comprising the training set, and 630 query images, which formed the test set. Image retrieval performance was evaluated by two frequently employed metrics, mean average precision at k (mAP@k) and precision at k (P@k).

B. Experiment of Feature Extraction Networks

As shown in Table II, TKANet demonstrates superior performance compared to representative lightweight networks, achieving higher CBIR accuracy while exhibiting more than 2x faster feature extraction speed. In Stage 1, CBIR acts as a coarse-grained filtration process, narrowing down candidate images for subsequent fine-grained filtration. The high-precision CBIR retrieval result attained by TKANet establishes a strong foundation for the subsequent LCBIR in Stage 2.

C. Experiment of Land Cover Classification Networks

As presented in Table III, our ESFormer surpasses competitive methods by a considerable margin, supported by the evidence presented in Figure 4. In this figure, ESFormer demonstrates more precise land cover classification, as evidenced by the accurate boundary reconstruction of the buildings, cars, trees, and pavement. The superior land cover classification results achieved by ESFormer offer more precise statistical information about land cover. This enhanced data contributes to refining the candidate images shortlisted by CBIR in Stage 1, thereby improving the effectiveness of LCBIR in Stage 2.

TABLE III
STAGE 2 LAND COVER CLASSIFICATION RESULTS OF VARIOUS METHODS

Method	MIoU (%)
FCN	51.01
PSPNet	54.69
UNet	61.17
Unet3+	61.29
UNet++	61.39
MA-UNet	61.90
GCN	62.85
DeepLabV3+	63.92
MACU	63.99
CANet	64.09
DI-Net	64.34
DDCM	64.43
SBANet	64.83
FE-Net	65.31
BAMTL	65.55
UNet+DC&LFC	65.64
FSCNet	67.31
Segformer	72.97
FURSformer	75.32
ESFormer	**76.19**

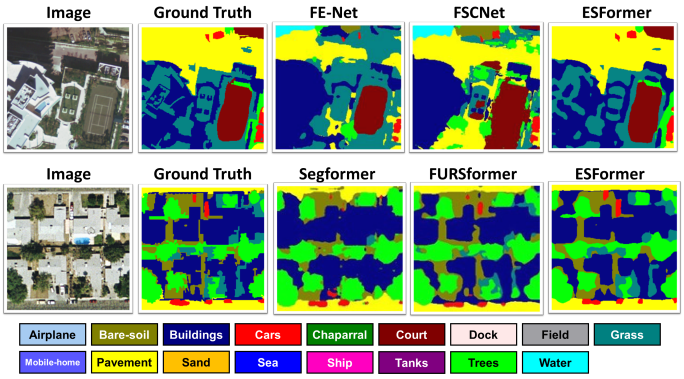

Fig. 4. Comparison of the predictions by various methods on DLRSD dataset.

D. Experiment of Image Retrieval Methods

We designated K, denoting the number of shortlisted images in HIR Stage 1, as 75. As detailed in Table IV, across diverse network combinations, HIR consistently enhances retrieval accuracy across the four evaluation metrics when compared to single-stage image retrieval methods.

Since we withholded the "Dense Residential", "Medium Residential", and "Sparse Residential" subcategory labels from the networks during the training process, to assess whether Stage 2 of HIR could refine Stage 1 retrieval outcomes to images sharing the same residential density as the query image, we calculated the retrieval accuracy of the broader category "Residential" exclusively. We assigned the weight w of the land cover class "Buildings" as 4 to emphasize residential density discrimination. Although fine-tuning the

Fig. 5. Comparison of "Medium Residential" Image Retrieval. Labels for residential sub-categories displayed above each image were withheld from networks.

TABLE IV
ALL CLASSES IMAGE RETRIEVAL RESULTS

Method	mAP@10	mAP@20	P@10	P@20
ResNet18 + Segformer				
LCBIR	0.8303	0.8035	0.7617	0.7209
CBIR	0.9478	0.9360	0.9246	0.9105
HIR (K=75)	**0.9591**	**0.9482**	**0.9377**	**0.9234**
TKANet + ESFormer				
LCBIR	0.8649	0.8381	0.7987	0.7510
CBIR	0.9532	0.9502	0.9457	0.9434
HIR (K=75)	**0.9623**	**0.9618**	**0.9591**	**0.9543**

TABLE V
RESIDENTIAL SUBCLASSES RETRIEVAL RESULTS BASED ON TKANET

Method	mAP@10	mAP@20	P@10	P@20
CBIR	0.6551	0.6191	0.5867	0.5267
LCBIR	0.7646	0.7127	0.6387	0.5273
HIR (K=75, w=1)	0.8454	0.7998	0.7627	0.6753
HIR (K=75, w=4)	**0.8501**	**0.8166**	**0.7880**	**0.7227**

hyperparameters K and w typically leads to improved results. However, to ensure generality, we refrained from fine-tuning the hyperparameters in our experiments. As presented in Table V, our HIR demonstrates clear superiority over conventional single-stage image retrieval methods. This substantiates HIR's capability to distinguish implicit sub-categories, thereby enhancing image retrieval for improved similarity. As evidenced in Figure 5, HIR and Weighted HIR gather an obviously higher percentage of "Medium Residential" images compared to CBIR and LCBIR. Notably, all images collected by HIR and Weighted HIR fall under the broader category "Residential". In contrast, both CBIR and LCBIR retrieved a few images categorized as "Storage Tank" and "Intersection", which differ from the query image.

While our two-stage method may introduce errors at each stage, Stage 2 leverages pixel class statistics to exclude images with dissimilar land cover class distributions from the query image. Despite land cover classification networks achieving around 70% accuracy on the mIoU metric, they surpass 90% accuracy on the pixel accuracy metric, which directly correlates with pixel class statistics. Given the high accuracy of pixel class statistics, errors in Stage 2 have a negligible influence on the final retrieval results. In the case of the DLRSD dataset, ESFormer achieves an inference speed of 153 frames per second on an NVIDIA RTX3050 Mobile GPU, with Stage 2 processing time of approximately 0.5 second.

IV. CONCLUSION

This study introduces a Hierarchical Image Retrieval (HIR) approach that integrates visual and semantic similarities. The experimental results indicate that HIR achieves 20% higher retrieval accuracy for "residential" sub-categories and over 1% increase in retrieval accuracy across all classes. In our future research, we aim to enhance HIR by substituting CBIR with cross-modal image retrieval methods in Stage 1.

REFERENCES

[1] Q. Zhu, Y. Sun, Q. Guan, L. Wang, and W. Lin, "A weakly pseudo-supervised decorrelated subdomain adaptation framework for cross-domain land-use classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[2] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, 2022.

[3] Q. Zhu, Y. Lei, X. Sun, Q. Guan, Y. Zhong, L. Zhang, and D. Li, "Knowledge-guided land pattern depiction for urban land use mapping: A case study of chinese cities," *Remote Sensing of Environment*, vol. 272, p. 112916, 2022.

[4] W. Zhou, H. Guan, Z. Li, Z. Shao, and M. R. Delavar, "Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 1447–1473, 2023.

[5] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sensing*, vol. 10, no. 6, p. 964, 2018.