Unsupervised Transformer Balanced Hashing for Multispectral Remote Sensing Image Retrieval

Yaxiong Chen, Fan Wang, Lin Lu, and Shengwu Xiong

***Abstract*—For remote sensing (RS) image retrieval task, hashing technology have been extensively researched in recent works. Unsupervised hashing approaches have attracted much attention in the RS data processing field because label collection takes a lot of time. Most of which fail to consider the interactions among the multichannel information of multispectral RS images and the disparity between the hash-like codes space and the Hamming space, which lead to the poor performance of multispectral RS image retrieval. In this article, we tackle these dilemmas with a novel unsupervised hashing approach, namely *Unsupervised Transformer Balanced Hashing* (UTBH), to utilize a convolutional variational autoencoder architecture with a novel RS transformer to perform effective hash codes learning. We first integrate a convolutional variational autoencoder architecture with a novel RS transformer, which can guide the interactions among the multichannel information of multispectral RS images. Meanwhile, a new objective function is proposed to preserve discrimination of hash codes in the hashing learning process and reduce the disparity between the hash-like codes space and the Hamming space effectively. Finally, experimental results on two multispectral RS image datasets indicate that UTBH approach achieves superior performance over other unsupervised image retrieval approaches.**

***Index Terms*—Hash codes, multichannel information, transformer, variational autoencoder.**

I. Introduction

With the high progress of remote sensing (RS) technology, RS images have shown a high-speed growth trend [1], [2]. Unearthing serviceable information from large-scale RS images is very critical [3], [4]. Hence, many researchers pay attention to the research of remote sensing image retrieval (RSIR) because RSIR can quickly find effective information from large-scale RS images [5], [6]. The goal of RSIR technology is to automatically match the RS image with similar semantics to the query RS image. Many content-based RSIR approaches have been developed gradually to manage and analyze RS images [7], [8], [9]. However, with the advancement of RS image acquisition equipment, early content-based RSIR approaches face the problems of slow retrieval speed and insufficient storage.

Hash technology is widely exploited to solve the problems of content-based RSIR approaches due to its fast speed and small storage space. The aim of hash technology is to map RS image into hash codes while conserving the similarity for RS image [10]. Existing hashing technology is divided into supervised category and unsupervised category. Supervised hashing algorithms learn hash function by leveraging supervised information. For example, Li et al. [6] introduced a new large-scale RS image retrieval algorithm, which leveraged labeled information to learn deep hash function. However, it is time-consuming to obtain these class labels. To solve the problem, unsupervised hash algorithms are widely used in large-scale RS image retrieval because they do not need the label information of datasets. For example, Wang et al. [11] proposed an unsupervised variational autoencoder hash algorithm, which can exploit multichannel feature fusion to learn hash codes for multispectral RS image retrieval.

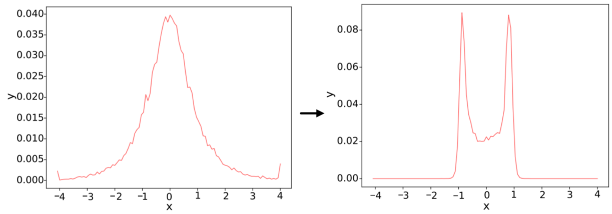
Despite existing unsupervised image retrieval algorithms have made some progress [5], [11], [12], there are still two obvious shortcomings. On the one hand, existing methods [5], [11] fail to reduce the disparity between the hash-like codes space and the Hamming space adequately, which ultimately leads to the poor multispectral image retrieval performance. On the other hand, most of the existing methods [11], [12] cannot consider the interactions among the multichannel information of multispectral RS images, which will lead to the issue of insufficient exploitation of multichannel information and eventually affects the experimental result.

In this article, we propose a novel multispectral RS image hashing approach, termed *Unsupervised Transformer Balanced*

Manuscript received 21 April 2023; revised 4 June 2023 and 28 June 2023; accepted 12 July 2023. Date of publication 26 July 2023; date of current version 7 August 2023. This work was supported in part by the NSFC under Grant 62101393, in part by the Project of Sanya Yazhou Bay Science and Technology City under Grant SCKJ-JYRC-2022-76 and Grant SCKJ-JYRC-2022-17, in part by the National Key Research and Development Program of China under Grant 2022ZD0160604, in part by the NSFC under Grant 62176194, in part by the Major Project of IoV under Grant 2020AAA001, in part by the Youth Fund Project of Natural Science Foundation of Hainan Province under Grant 6220N344, in part by the CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2022-001A, in part by the Knowledge Innovation Program of Wuhan-Basic Research, in part by the Sanya Science and Education Innovation Park of Wuhan University of Technology under Grant 2022KF0020, in part by the Fundamental Research Funds for the Central Universities under Grant WUT:223110001, in part by High-performance Computing Platform of YZBSTCACC, in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX1148, in part by the Hainan Special PhD Scientific Research Foundation of Sanya Yazhou Bay Science and Technology City under Grant HSPHDSRF-2023-03-002, and in part by MindSpore, which is a new deep learning computing framework (https://www.mindspore.cn/). *(Corresponding author: Shengwu Xiong.)*

The authors are with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China, also with the Shanghai Artificial Intelligence Laboratory, Wuhan University of Technology, Shanghai 200232, China, also with the Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China, and also with the Chongqing Research Institute, Wuhan University of Technology, Chongqing 401122, China (e-mail: 593544199@qq.com; wangfans@whut.edu.cn; linklu@whut.edu.cn; xiongsw@whut.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3298942



(a) (b)

Fig. 1. Core idea of the proposed UTBH approach. (a) Distribution of network outputs for the proposed UTBH approach without using the balanced term. (b) Distribution of network outputs for the proposed UTBH approach. The balanced term can obviously reduce the disparity between the hash-like codes space and the Hamming space.

Hashing (UTBH), to conduct hash codes generation by guiding the interactions of the multichannel information for multispectral RS image as well as reducing the disparity between the hash-like codes space and the Hamming space, as demonstrated in Fig. 2. UTBH approach integrates a convolutional variational autoencoder architecture with a novel RS transformer, which can guide the interactions of the multichannel information of multispectral RS image. In addition, a novel objective function is developed by the composite of reconstruction cost, KL divergence, and balanced term, which conserve the discrimination of hash codes in the hashing learning process and reduce the disparity between the hash-like codes space and the Hamming space. Fig. 1(a) demonstrates the distribution of network outputs for UTBH approach without using the balanced term. Fig. 1(b) demonstrates the distribution of network outputs for UTBH approach. It can be seen from Fig. 1(a) and (b) that the balanced term can obviously reduce the disparity between the hash-like codes space and the Hamming space. Abundant experiments on diverse benchmark have well demonstrated the reasonableness and effectiveness of the proposed UTBH approach.

The contributions can be summarized as follows:

1) A novel unsupervised RSIR framework is developed to leverage the transformer for solving the issue of deficient usage of the interactions among the multichannel information of multispectral RS images. As far as we know, it is the first work to conduct hash codes generation by considering the interactions among the multichannel information of multispectral RS images.
2) Since the disparity between the hash-like codes space and the Hamming space is inescapable, a novel objective function is proposed to reduce the disparity between the hash-like codes space and the Hamming space effectively.
3) Abundant experiments on diverse benchmark for multispectral RS images demonstrate that UTBH can learn better hash codes, which achieve more effective retrieval performance than retrieval approaches.

II. Related Works

In this section, according to whether deep learning is adopted, related works are divided into two aspects: traditional RSIR and deep learning-based RSIR.

A. Traditional RSIR

Traditional RSIR approaches exploit hand-crafted features to perform RSIR. For example, Zhu et al. [13] leveraged Gabor filter to extract image texture features, which can be exploited to retrieve aerial RS images by calculating Euclidean distance. Li et al. [14] utilized Gabor texture features to retrieve RS images. Scott et al. [15] automatically extracted object features of multiple scales from large-scale satellite RS images. Then, object features were applied to retrieve objects. Shao et al. [16] used visual attention model to obtain salient objects. The color features and texture features of salient objects were combined for image retrieval. Chaudhuri et al. [17] utilized the shape features and texture features of RS images to form a new feature vector. Then, the new feature vector was leveraged to realize multilabel RS image retrieval by using semisupervised way.

However, the retrieval effect of shape features and texture features is poor. For better retrieval, many researchers proposed many advanced hand-crafted features such as SIFT features, GIST features, and SURF features. For example, Douze et al. [18] verified the rationality and accuracy of GIST features and proposed a global feature index optimization strategy to balance memory utilization and retrieval accuracy. An et al. [19] developed SURF features to retrieve RS video, which improved the significance of image features and shortened the time of feature generation and matching. Newsam et al. [20] extracted the SIFT descriptor of RS images, which was leveraged to perform RSIR. Yang et al. [21] leveraged local invariant features to improve the RSIR performance. Zhou et al. [22] proposed a RSIR approach, which leveraged sparse representation theory and the topology information of features.

Many existing studies on traditional RSIR methods focus on using hand-crafted features to perform RSIR. Different from these methods, the proposed UTBH method leveraged deep learning framework to perform RSIR.

B. Deep Learning-Based RSIR

With the rapid development of satellite and aircraft technology, the era of RS Big Data has come [23]. Unearthing serviceable information from large-scale RS images is very important. The retrieval performance of traditional RSIR mainly depends on the sensor type, band information, and geographical location of manually labeled RS images. These methods take a lot of time and cannot accurately reflect the high-level semantic information of RS images.

To solve these problems, deep learning-based RSIR are proposed to utilize deep convolution neural network to learn high-level semantic information of RS images, which can significantly improve the accuracy and speed of RSIR. For example, Ye et al. [24] leveraged high-level semantic information and SIFT features of RS images to construct feature sets to match RS images. Liu et al. [25] transformed similarity learning into deep ordered classification learning to tackle the issue of relying on a large number of labeled samples in traditional RSIR. Kumar et al. [26] utilized CNN to extract high-level semantic features of RS building images for RSIR. Imbriaco et al. [27] added saliency

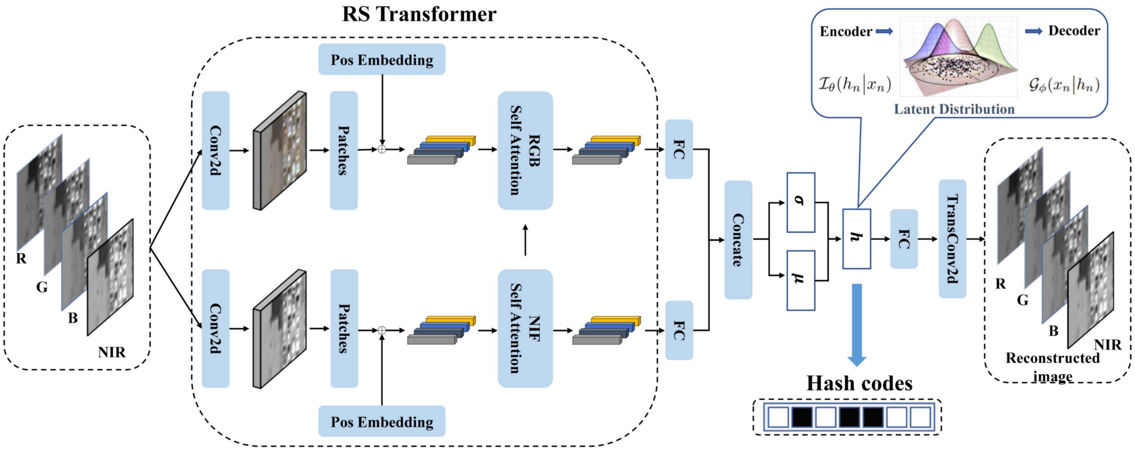


Fig. 2. Overview of the proposed UTBH framework. The input of UTBH model is multispectral RS image, which contains red (R), green (G), blue (B), and near infrared (NIR) channel. The framework integrates a convolutional variational autoencoders architecture with a RS transformer. The overall objective function consists of reconstruction cost, KL divergence, and balanced term.

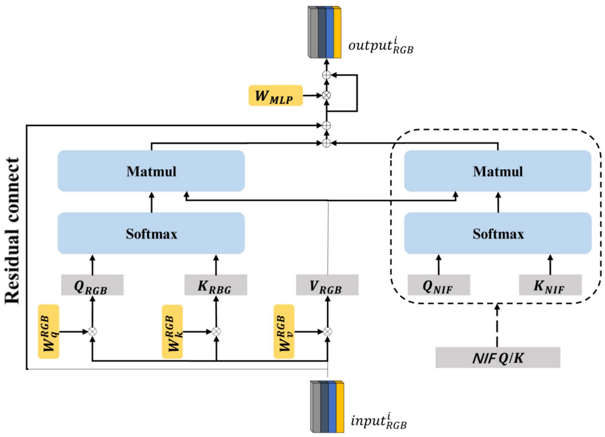


Fig. 3. Multihead self-attention process of NIF self-attention module.

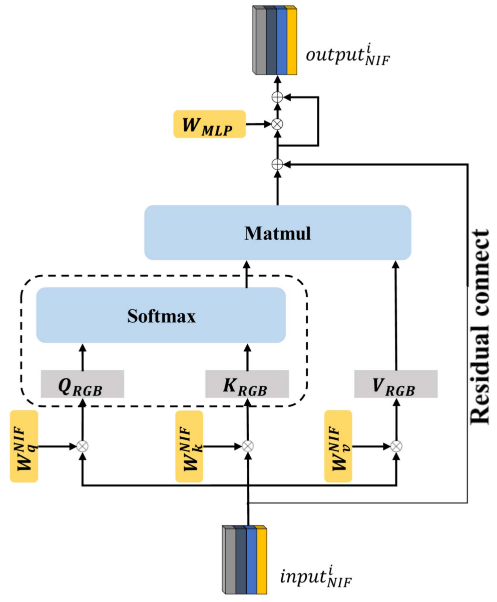


Fig. 4. Multihead self-attention process of RGB self-attention module.

module to obtain multiscale convolution aggregation features to perform RSIR.

Although the effect of deep learning-based RS image retrieval is good, massive data have higher requirements for storage space and computational complexity. Deep hashing technology is widely applied to solve the problem of large-scale RSIR due to its low computational complexity and small storage space. For example, Zhu et al. [28] developed a novel hash codes learning approach with multiple features learning for RS images. Li et al. [5] introduced a novel partial randomness hashing method, which mapped RS images to low-dimensional feature representation by random projection in an unsupervised manner for large-scale RSIR. Reato et al. [29] introduced an unsupervised hash codes learning method for accurate and scalable RSIR. Kang et al. [30] proposed a new deep-hashing technique based on the class-discriminated neighborhood embedding, which can properly capture the locality structures among the RS scenes and distinguish images classwisely in the Hamming space. Han et al. [31] developed a deep cohesion intensive network, which not only improved the retrieval performance of RS images, but also overcame the imbalance problem of RS images.

As far as we know, few works perform unsupervised RS image retrieval by using hash codes [11], [12]. However, existing deep learning-based RSIR methods cannot consider the interactions among the multichannel information of multispectral RS images and reduce the disparity between the hash-like codes space and the Hamming space, which affect the retrieval performance. To improve the retrieval performance, UTBH can guide the interactions among multichannel information of multispectral RS images and reduce the disparity between the hash-like codes space and the Hamming space effectively.

III. Proposed Method

In this section, we first clarify the conception of RSIR in Section III-A. Second, Section III-B presents the architecture of UTBH. Third, Section III-C introduces the details of RS transformer. Finally, Section III-D provides the objective function of

UTBH. The deep architecture of the proposed UTBH approach, which can guide the interactions between multichannel information of multispectral RS images and reduce the disparity between the hash-like codes space and the Hamming space effectively, as demonstrated in Fig. 2.

A. Notation

Given N RS images $\boldsymbol{\mathcal{X}} = \{x_n\}_{n=1}^{N}$, where x_n denotes the nth RS image. The aim of hash codes learning is to generate hash function $\mathcal{H}$ that projects RS image x_n into hash codes $b_n \in \{-1, 1\}^k$ while preserving the similarity of RS image [32], [33], [34]. Existing RS image retrieval approaches ignore the interactions among the multichannel information of multispectral RS images. Hence, we develop a RSIR approach called UTBH to perform hash codes learning by guiding the interactions among the multichannel information of multispectral RS images and reducing the disparity between the hash-like codes space and the Hamming space. As demonstrated in Fig. 2, the UTBH approach consist of two components: 1) The proposed UTBH model integrates a convolutional variational autoencoders architecture with a novel RS transformer, which can guide the interactions among the multichannel information of multispectral RS images. 2) A novel objective function is developed by the composite of reconstruction cost, KL divergence and balanced term, which capture discrimination of hash codes in the hash codes learning process and reduce the disparity between the hash-like codes space and the Hamming space. In the following section, we first describe the architecture of the proposed UTBH approach, and then we present the details of RS transformer. Finally, we clarify the objective function.

B. Model Architecture

The proposed framework of UTBH is demonstrated in Fig. 2. The proposed framework leverages variational autoencoder as the backbone network, which contains the inference network and the generation network. The details are explained as follows.

Inference network: The inference network is exploited to implement variational inference on the raw data and produces the variational probability distribution of latent variables. Specifically, the inference network $\mathcal{I}_\theta(h_n|x_n)$ projects the raw data x_n into the variational probability distribution. Then, the feature vector h_n is sampled from the variational probability distribution. The inference network consists of RS transformer, two parallel fully connected layer and a hashing encoding layer. The parallel fully connected layer contains k nodes. The hashing encoding layer leverages the reparameterization trick to connect two parallel fully connected layer. The details of RS transformer are introduced in Section III-C.

Generation network: The generation network is leveraged to restore the approximate probability distribution of the raw data from the variational probability distribution of latent variables. Specifically, generation network $\mathcal{G}_\phi(x_n|h_n)$ projects feature vector h_n to reconstruct x_n. The generation network consist of a fully connected layer, a reshape operator, four transposed convolution layers with BN [35], and a convolutional layer [36], [37], [38]. Four transposed convolution layers exploit 256, 128, 64, and 32 filters with size 3×3. The stride of transposed convolution layers is 2 pixels. Four transposed convolution layers utilize the *LeakyReLU function* as the activation function. A convolutional layer exploits three filters with size 3×3. The stride of convolution layer is 1 pixel. It utilizes the *tanh function* as the activation function.

Accordingly, given any RS image x_n, deep hashing function is formulated as

$$b_n = \mathcal{H}(x_n) = \text{sign}(\mathcal{I}_\theta(h_n|x_n)) \tag{1}$$

with

$$\text{sign}(x) = \begin{cases} 1, & x \geqslant 0 \\ -1. & x < 0 \end{cases} \tag{2}$$

where b_n denotes k-bits hash codes for instance x_n, $\mathcal{H}$ denotes deep hashing function for instance x_n. θ denotes the parameters of the inference network.

C. RS Transformer

Following [39], [40], [41], RS transformer consists of a convolution layer, a patch reshape operator, a position embeddings, RGB self-attention module, and NIF self-attention module. The convolution layer exploit 128 filters with size 8×8. The stride of the convolution layer is 8 pixel. The patch reshape operator can reshape the feature maps to 16 patch embeddings with 128 dims. Position embeddings can be added to the patch embeddings. Specifically, 17 standard learnable 1-D position embeddings with 128 dims are leveraged to retain positional information and the site of 0th position embedding is added to a standard token, which contains 128-D learnable parameter. The resulting sequence of embedding vectors serves as input to the RGB self-attention module and NIF self-attention module.

NIF self-attention module: The NIF self-attention module employed six alternating stacks of eight multiheaded self-attention and MLP layers. In other words, one alternating stack consists of two sublayer, which applied a layer norm before each sublayer and residual connection after each sublayer. So the output of each sublayer can be formulated as

$$\text{SubLayerOut} = \text{LayerNorm}(I_n + (\text{SubLayer}(I_n))) \tag{3}$$

where $SubLayerOut$ denotes the output of each sublayer. I_n denotes the sublayer input.

Multihead self-attention process of NIF self-attention block can be shown in Fig. 3. input_{NIF}^{i} denotes the input patch embeddings of stack i. output_{NIF}^{i} is served as input of the $i+1$ stack input_{NIF}^{i+1}. We employ these parameter weights W_q^{NIF}, W_k^{NIF}, W_v^{NIF} to transfer input_{NIF}^{i} to matrices Q_{NIF}^{i}, K_{NIF}^{i}, V_{NIF} and the attention sublayer is formulated as follows:

$$\text{attention}_{NIF} = \text{softmax}\left(\frac{Q_{NIF}^{i} {K_{NIF}^{i}}^{T}}{\sqrt{d_k}}\right) V_{NIF}^{i} \tag{4}$$

where softmax$(\cdot)$ denotes the softmax function. To take into account the speed and space-efficient [42], the dot products can be scaled by $\sqrt{d_k}$.

RGB self-attention module: The RGB module consists a stack of six identical layers. Unlike the NIF attention sublayer in each

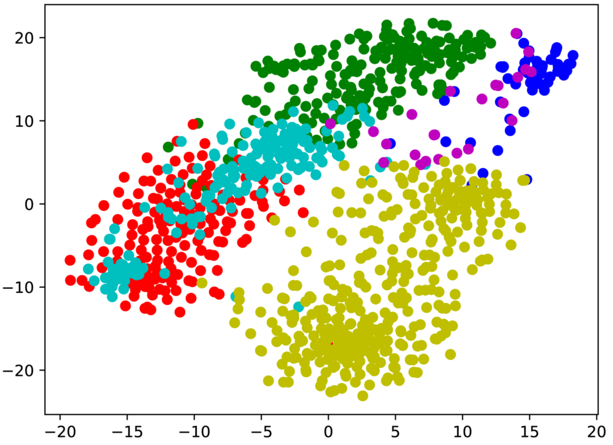


Fig. 5. t-SNE visualizations of the hash-like codes of the proposed UTBH-L approach without considering the balanced term on SAT-6 dataset.

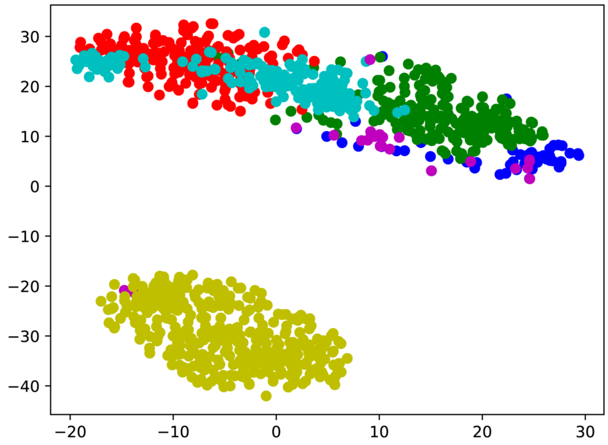


Fig. 6. t-SNE visualizations of the hash-like codes of the proposed UTBH approach on SAT-6 dataset.

stack, the RGB self-attention sublayer inserts a second mask function. Multihead self-attention process of RGB self-attention module is shown in Fig. 4. Similar to NIF attention layer, the RGB attention sublayer is formulated as follows:

$$\text{attention}_{RGB} = \text{softmax}\left(\frac{Q^i_{RGB}{K^i_{RGB}}^T}{\sqrt{d_k}}\right)V^i_{RGB} + \text{softmax}\left(\frac{Q^i_{NIF}{K^i_{NIF}}^T}{\sqrt{d_k}}\right)V^i_{RGB} \quad (5)$$

where Q^i_{RGB} denotes the query matrices of RGB self-attention module. $\{K^i_{RGB}, V_{RGB}\}$ denote paired key-value matrices of RGB self-attention module.

D. Objective Function

Fig. 2 shows the brief overview of the proposed UTBH approach. To generate effective hash codes, reconstruction cost can be given as

$$\mathcal{J}_{r1} = \sum_{i=1}^{N} \|x_n - \mathcal{G}_\phi(x_n|b_n)\|^2 \quad (6)$$

where $\mathcal{G}_\phi(x_n|b_n)$ denotes the generation network of b_n. The reconstruction cost is leveraged to constrain the input value before coding and the output value after decoding, so that the reconstructed data of variational autoencoder is still similar to the original data. However, $\mathcal{J}_{r1}$ in (6) is a nonsmooth function, which is hard to calculate the derivative in the deep neural network training process. Following [43], reconstruction cost can be rewritten as

$$\mathcal{J}_{r} = \sum_{i=1}^{N} \|x_n - \mathcal{G}_\phi(x_n|h_n)\|^2 \quad (7)$$

where $\mathcal{J}_r$ denotes reconstruction cost and h_n denotes hash-like codes.

In addition, the variational autoencoder needs to keep the variational probability distribution approximate the standard normal distribution $N(0,1)$ by minimizing KL divergence. The KL divergence can be defined as

$$\mathcal{J}_{u,\sigma^2} = KL(N(\mu_i, {\sigma_i}^2))\|N(0,1)) \quad (8)$$

where $\mathcal{J}_{u,\sigma^2}$ denotes the KL divergence, which can preserve discrimination of hash codes in the learning process.

Since the disparity between the hash-like codes space and the Hamming space is inescapable, a novel balanced term is proposed to reduce the disparity between the hash-like codes space and the Hamming space effectively. Following [44], [45], the balance rule expects each value of the hash-like codes to approximate -1 or 1. To conserve the balance property of hash codes, the balanced term can be defined as

$$\mathcal{J}_b = \begin{cases} 0, & |\mu_i| \geqslant 1 \\ \sum_{i=1}^{N}(1-\mu_i^2). & -1 < \mu_i < 1 \end{cases} \quad (9)$$

where $\mathcal{J}_b$ denotes the balanced term, which can reduce the disparity between the hash-like codes space and the Hamming space effectively. The balanced term can obviously encourage each value of the network outputs of hash-like codes to approximate -1 or 1.

By considering the above three parts (i.e., reconstruction cost $\mathcal{J}_r$, KL divergence $\mathcal{J}_{u,\sigma^2}$, and balanced term $\mathcal{J}_b$), the total objective function can be formulated as

$$\mathcal{J} = \mathcal{J}_r + \alpha\mathcal{J}_{u,\sigma^2} + \beta\mathcal{J}_b \quad (10)$$

where α and β denote the hyperparameters that evaluate the degree of term. $\mathcal{J}$ denotes the total objective function, which can preserve discrimination of hash codes in the hashing learning process and reduce the disparity between the hash-like codes space and the Hamming space effectively.

IV. Experiments

In this section, we clarify two RS datasets and evaluation protocols in Section IV-A. The implementation details of UTBH in Section IV-B. Section IV-C introduces different factors of UTBH. Section IV-D analyzes the experiment of UTBH. Section V presents the conclusion of UTBH.

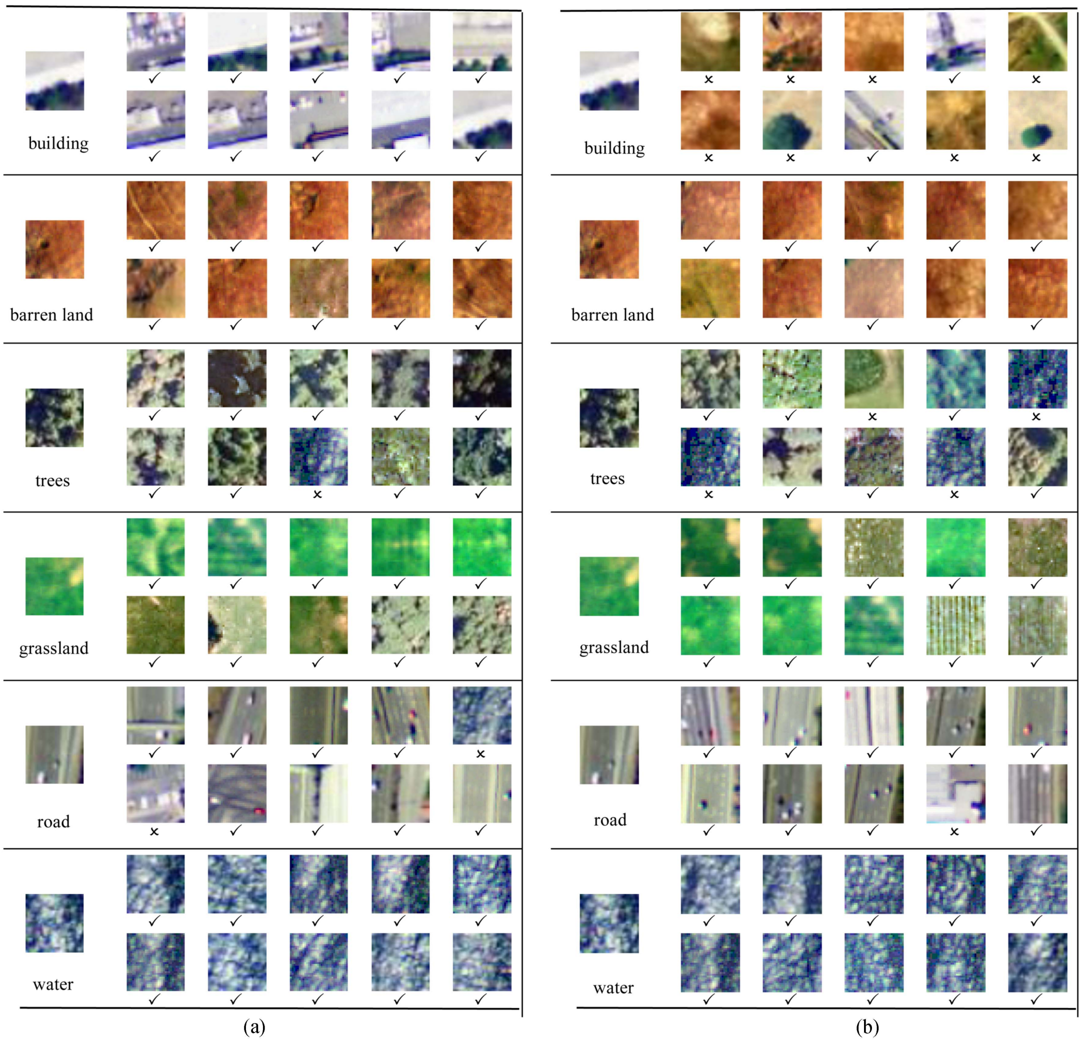


Fig. 7. Top 10 retrieval samples with 64 bits on SAT-6 dataset. (a) UTBH. (b) UTBH-T. The false retrieval samples are marked with a cross. The right retrieval samples are marked with a tick.

Algorithm 1: UTBH Algorithm.

```
Input:
  $N$ trained samples $\{x_i\}_{i=1}^{N}$.
Output:
  The parameters $\theta$ and $\phi$ of UTBH approach.
Initialization:
  The parameters $\theta$ and $\phi$ are initialized by glorot _
    uniform distribution.
Repeat:
  Utilize generation network $\mathcal{I}_\theta(h_n|x_n)$ to obtain
    hash-like codes $h_n$;
  Leverage generation network $\mathcal{G}_\phi(x_n|h_n)$ to
    reconstruct $x_n$;
  Compute hash codes $b_n$ according to (1);
  Compute the objective function $\mathcal{J}$ by (10);
  Update the parameters $\theta$ and $\phi$ by utilizing Adam.
Until: a fixed number of iterations.
Return: $\theta$ and $\phi$.
```

A. Dataset and Evaluation Protocols

To demonstrate the effectiveness of UTBH, three public RSI datasets are leveraged to compare the proposed UTBH approach with other RSIR approaches.

1) SAT-4 dataset contains 500 000 multispectral RS images [46]. Each multispectral RS image has four channels, which represent red (R), green (G), blue (B), and near infrared (NIR). The size of each multispectral RS image is $28 * 28$. Following [11], the proposed UTBH approach randomly conduct 1000 multispectral RS images to constitute the test and retrieval set. Moreover, the remainder multispectral RS images can be selected to constitute the training set.

2) SAT-6 dataset contains 405 000 multispectral RS images [46]. Each image has four channels. The size of multispectral RS image is 28×28. The SAT-6 dataset has six categories. Following [11], the proposed UTBH approach randomly selects 1000 RS images to constitute

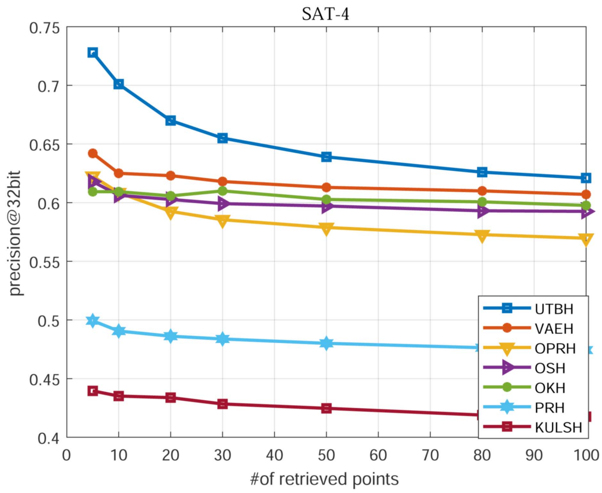


Fig. 8. Precision curves with different returned samples with 32 bits in the SAT-4 dataset.

the test and retrieval set. The remainder RS images can be selected to constitute the training set.

3) The AID dataset [47] contains 10 000 images with 30 classes. The size of RS image is 600 × 600. RGB channels are used to generate NIR [48]. Then, we use NIR and RGB to form a multispectral image. In the training stage, 20% images are randomly utilized as a test query set, and the rest 80% images are regarded as the training. These evaluating metrics mAP and *the precision in top-m of the ranking list* are leveraged to evaluate experimental results of the proposed UTBH approach [33], [43], [49], [50].

B. Implementation Detail

The proposed UTBH approach is carried out by leveraging the open-source Pytorch[1] and Mindspore[2] library. The experiments are carried out on workstation with Inter Core i7−5930 K 3.50 GHZ CPU, 64 G RAM, and GeForce GTX Titan X GPU. The proposed UTBH network is optimized by exploiting Adam [51]. The batch size of the proposed UTBH network is fixed as 512. To produce {32, 48, 64}-bit hash codes, hash code length k is fixed from 32 to 64, individually. The weight initialization of UTBH exploit glorot _ uniform distribution. The parameter α is fixed to 1. β is fixed to 5. The proposed UTBH network is trained for five epochs, or stop training until the loss does not decline [52].

C. Evaluation of Different Parts

To estimate the significance of the RS transformer and the balanced term in the UTBH model, the experiments are carried out at the following points: First, we exploit the UTBH model without the RS transformer to learn hash function (i.e., UTBH-T). Second, we utilize the UTBH model without leveraging the balanced term to perform cross-modal learning (i.e., UTBH-L). Finally, we implement the proposed approach (i.e., UTBH).

For the SAT-4 dataset, Table I presents the comparison of mean precision of top-10 ranked results and top-100 ranked results for UTBH-T, UTBH-L, and UTBH with 32, 48, and 64 bits. For the SAT-6 dataset, Table II presents the comparison of mean precision of the top-10 ranked results and the top-100 ranked results for UTBH-T, UTBH-L, and UTBH with 32, 48, and 64 bits. It can be seen from Tables I and II that UTBH obtains better performance than UTBH-T and UTBH-L. For example, for the top-10 ranked results with 32 bits, UTBH can enhance the mAP to 70.1% from 62.5% carried out by UTBH-T, 64.3% carried out by UTBH-L. For the top-100 ranked results with 32 bits, UTBH can enhance the mAP to 62.1% from 52.2% carried out by UTBH-T, 55.4% carried out by UTBH-L. This is because UTBH exploits the RS transformer and the balanced term to learn more efficient hash codes. Specially, Fig. 5 presents t-SNE visualizations of hash-like codes of the proposed UTBH approach without considering the balanced term on SAT-6 dataset. Fig. 6 shows t-SNE visualizations of hash-like codes of the proposed UTBH approach on SAT-6 dataset. As demonstrated in Fig. 5, UTBH without exploiting the balanced term fails to generate the discriminative hash codes. Because it cannot reduce the disparity between the hash-like codes space and the Hamming space. Besides, we demonstrate the comparability between UTBH and UTBH-T in Fig. 7, where top ten ranked RS images with 64 bits on SAT-6 dataset obviously speculate the significance of UTBH retrieval result.

TABLE I
COMPARISON OF MEAN PRECISION OF THE TOP-10 RETRIEVED RESULTS AND THE TOP-100 RETRIEVED RESULTS FOR UTBH-T, UTBH-L AND UTBH ON THE SAT-4 DATASET WITH DIFFERENT HASH BITS

Task	Constraint	32bits	48bits	64bits
Top-10	UTBH-T	62.5	66.7	69.3
	UTBH-L	64.3	68.1	71.4
	UTBH	70.1	70.7	72.0
Top-100	UTBH-T	52.2	57.0	61.0
	UTBH-L	55.4	58.3	62.7
	UTBH	62.1	62.2	63.2

TABLE II
COMPARISON OF MEAN PRECISION OF THE TOP-10 RETRIEVED RESULTS AND THE TOP-100 RETRIEVED RESULTS FOR UTBH-T, UTBH-L, AND UTBH ON THE SAT-6 DATASET WITH DIFFERENT HASH BITS

Task	Constraint	32bits	48bits	64bits
Top-10	UTBH-T	79.0	78.8	79.1
	UTBH-L	79.6	80.2	79.3
	UTBH	82.4	83.6	83.6
Top-100	UTBH-T	71.8	72.2	71.6
	UTBH-L	73.3	74.1	73.4
	UTBH	76.8	78.1	77.3

[1] [Online]. Available: https://pytorch.org
[2] [Online]. Available: https://www.mindspore.cn

TABLE III
COMPARISON OF MEAN PRECISION OF THE TOP-10 RETRIEVED RESULTS FOR DIFFERENT METHODS ON THE SAT-4 DATASET WITH DIFFERENT HASH BITS

	Method	32 bits(%)	48 bits(%)	64 bits(%)
Top-10	IMH [53]	56.0	53.8	54.8
	IsoHash [54]	60.6	64.0	65.5
	ITQ [55]	63.6	65.3	66.2
	SpH [56]	59.6	62.3	65.8
	KULSH [57]	49.2	50.7	55.3
	PRH [5]	60.7	62.1	66.5
	OKH [58]	43.9	51.6	60.0
	OSH [59]	60.3	63.7	64.7
	OPRH [12]	60.8	63.0	65.6
	VAEH [11]	62.5	65.0	66.6
	UTBH	**70.1**	**70.7**	**72.0**

The bold values represent the best results on the corresponding dataset.

TABLE IV
COMPARISON OF MEAN PRECISION OF THE TOP-100 RETRIEVED RESULTS FOR DIFFERENT METHODS ON THE SAT-4 DATASET WITH DIFFERENT HASH BITS

	Method	32 bits(%)	48 bits(%)	64 bits(%)
Top-100	IMH [53]	55.0	52.4	54.1
	IsoHash [54]	57.6	59.4	59.7
	ITQ [55]	60.9	60.7	61.0
	SpH [56]	56.3	58.8	60.7
	KULSH [57]	47.6	47.9	52.6
	PRH [5]	59.2	59.5	62.2
	OKH [58]	41.8	48.0	56.1
	OSH [59]	56.8	59.6	59.6
	OPRH [12]	59.8	59.4	61.6
	VAEH [11]	60.7	61.6	62.3
	UTBH	**62.1**	**62.2**	**63.2**

The bold values represent the best results on the corresponding dataset.

D. Method Comparison

1) Results on SAT-4 Dataset: For the SAT-4 dataset, to assess the significance of UTBH, we compare UTBH with several approaches, including IMH [53], IsoHash [54], ITQ [55], SpH [56], KULSH [57], PRH [5], OKH [58], OSH [59], OPRH [12], VAEH [11]. VAEH is a deep learning-based approach. IMH, IsoHash, ITQ, SpH, KULSH, PRH, OKH, OSH, and OPRH are traditional methods. Table III presents the comparison of mean precision of the top-10 ranked results for many approaches with 32, 48, and 64 bits. Table IV shows the comparison of mean precision of the top-100 ranked results for many approaches with 32, 48, and 64 bits. We can obviously notice that although comparative retrieval approaches have gained good outcome on two metrics, UTBH can attain the best precision in top ten ranked consequence and the best retrieval precision in top 100 ranked consequence on RS image dataset. For example, for the top-10 retrieved results, the proposed UTBH approach can enhance the mean precision with 64 bits from IMH (54.8%), IsoHash (65.5%), ITQ (66.2%), SpH (65.8%) KULSH (55.3%), PRH (66.5%), OKH (60.0%), OSH (64.7%), OPRH (65.6%), VAEH (66.6%) to 72.0%. Moreover, the top-100 retrieved results, UTBH can enhance the mean precision with 64 bits from IMH (54.1%), IsoHash (59.7%), ITQ (61.0%), SpH (60.7%) KULSH (52.6%), PRH (62.2%), OKH (56.1%), OSH (59.6%), OPRH (61.6%), VAEH (62.3%) to 63.2%. Fig. 8 shows precision curves with different returned samples with 32 bits. Fig. 9 presents precision curves with different returned samples with 48 bits. Fig. 10 presents precision curves with different returned samples with 64 bits. From these three figures, UTBH can gained better outcome than other comparative approaches. This is because UTBH can not only guide the interactions among the multichannel information of multispectral RS images, but also reduce the disparity between the hash-like codes space and the Hamming space adequately.

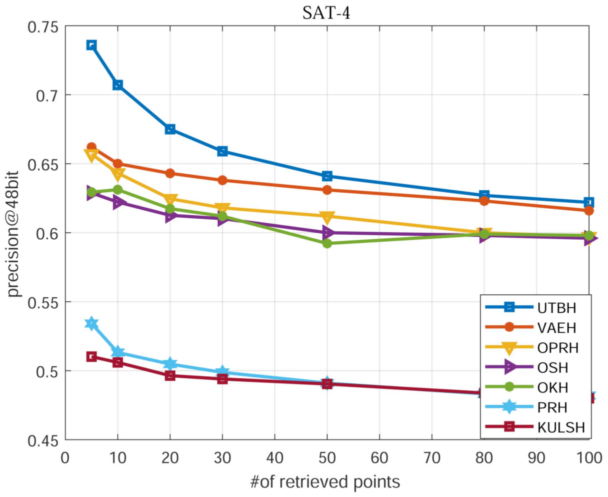


Fig. 9. Precision curves with different returned samples with 48 bits in the SAT-4 dataset.

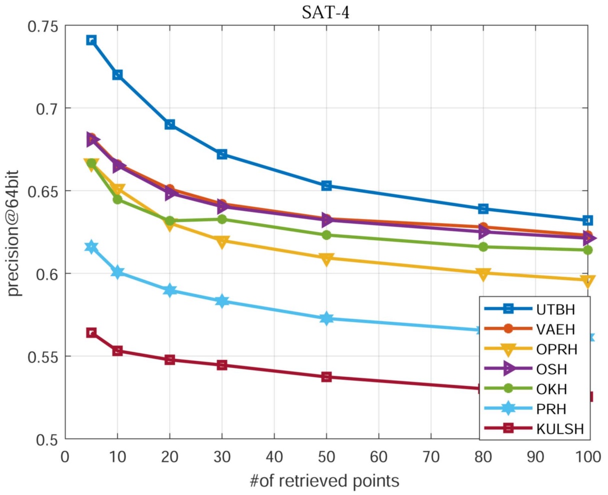


Fig. 10. Precision curves with different returned samples with 64 bits in the SAT-4 dataset.

2) Results on SAT-6 Dataset: For the SAT-4 dataset, the comparison of mean precision of the top-10 ranked results for many approaches with 32, 48, and 64 bits is shown in Table V. The comparison of mean precision of the top-100 ranked results for many approaches with 32, 48, and 64 bits is shown in Table VI. Fig. 11 presents precision curves for different ranked samples with 32 bits. Fig. 12 presents precision curves for different

TABLE V
COMPARISON OF MEAN PRECISION OF THE TOP-10 RETRIEVED RESULTS FOR DIFFERENT METHODS ON THE SAT-6 DATASET WITH DIFFERENT HASH BITS

	Method	32 bits(%)	48 bits(%)	64 bits(%)
Top-10	IMH [53]	58.3	62.6	60.4
	IsoHash [54]	66.7	68.0	67.3
	ITQ [55]	67.2	69.1	68.1
	SpH [56]	64.2	66.4	69.4
	KULSH [57]	41.3	45.9	45.2
	PRH [5]	65.1	68.2	68.3
	OKH [58]	54.1	61.9	63.8
	OSH [59]	66.9	68.4	68.0
	OPRH[12]	64.5	69.9	70.5
	VAEH[11]	69.1	73.2	74.8
	UTBH	**82.4**	**83.6**	**83.6**

The bold values represent the best results on the corresponding dataset.

TABLE VI
COMPARISON OF MEAN PRECISION OF THE TOP-100 RETRIEVED RESULTS FOR DIFFERENT METHODS ON THE SAT-6 DATASET WITH DIFFERENT HASH BITS

	Method	32 bits(%)	48 bits(%)	64 bits(%)
Top-100	IMH [53]	57.5	61.4	58.2
	IsoHash [54]	63.5	64.5	64.2
	ITQ [55]	64.9	66.0	65.3
	SpH [56]	61.6	63.1	65.7
	KULSH [57]	41.8	49.6	52.0
	PRH [5]	62.9	65.8	65.2
	OKH [58]	52.1	59.2	61.7
	OSH [59]	63.9	65.0	64.7
	OPRH [12]	63.1	67.2	67.7
	VAEH [11]	67.9	70.4	71.2
	UTBH	**79.5**	**78.1**	**77.3**

The bold values represent the best results on the corresponding dataset.

ranked samples with 48 bits. Fig. 13 presents precision curves for different ranked samples with 64 bits. Similar experimental results are clearly observed on SAT-4 image dataset. UTBH achieves the best performance on all the metrics, which illustrate the availability of generating hash codes by guiding the interactions among the multichannel information and reducing the disparity between the hash-like codes space and the Hamming space adequately.

3) Results on AID Dataset: To investigate the effectiveness of the proposed UTBH approach in RSI retrieval task, we compare UTBH with several approaches, including PRH [5], OSH [59], OPRH [12], GreedyHash [60], VAEH [11], and BihalfHash [61]. GreedyHash, VAEH, and BihalfHash are deep learning-based approaches. Table VII shows the comparison of mean precision of the top-10 ranked consequences and the top-100 ranked consequences for different algorithms on the AID dataset with different lengths. For example, for the top 10 retrieved results, the proposed UTBH approach can enhance the mean precision with 64 bits from PRH (15.78%), OSH (16.84%), OPRH (17.46%), GreedyHash (19.35%), VAEH

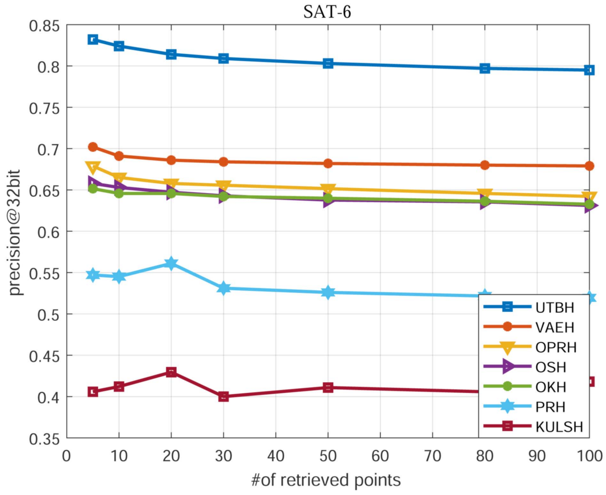


Fig. 11. Precision curves with different returned samples with 32 bits in the SAT-6 dataset.

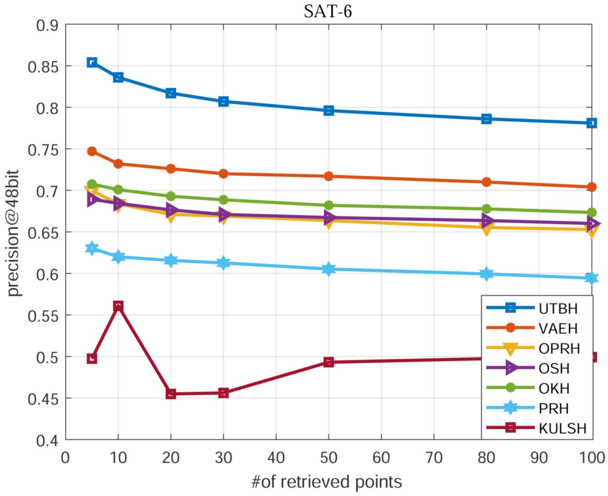


Fig. 12. Precision curves with different returned samples with 48 bits in the SAT-6 dataset.

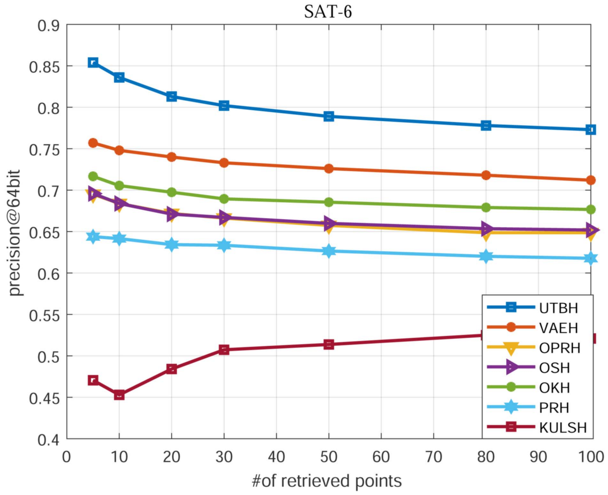


Fig. 13. Precision curves with different returned samples with 64 bits in the SAT-6 dataset.

TABLE VII
COMPARISON OF MEAN PRECISION OF THE TOP-10 RANKED CONSEQUENCES AND THE TOP-100 RANKED CONSEQUENCES FOR DIFFERENT ALGORITHMS ON THE AID DATASET WITH DIFFERENT LENGTHS

	Method	32 bits(%)	48 bits(%)	64 bits(%)
Top-10	PRH [5]	15.13	15.46	15.78
	OSH [59]	16.24	16.64	16.84
	OPRH [12]	16.68	16.94	17.46
	GreedyHash [60]	17.64	17.88	19.35
	VAEH [11]	18.41	19.28	20.39
	BihalfHash [61]	20.46	23.37	24.45
	UTBH	**22.14**	**23.72**	**25.63**
Top-100	PRH [5]	10.54	10.77	11.05
	OSH [59]	11.64	11.82	11.96
	OPRH [12]	11.95	12.16	12.63
	GreedyHash [60]	11.84	11.94	15.03
	VAEH [11]	11.84	13.27	14.74
	BihalfHash [61]	14.36	16.18	17.36
	UTBH	**15.14**	**16.26**	**17.86**

The bold values represent the best results on the corresponding dataset.

(20.39%), BihalfHash (24.45%) to 25.63%. Moreover, the top-100 retrieved results, UTBH can enhance the mean precision with 64 bits from PRH (11.05%), OSH (11.96%), OPRH (12.63%), GreedyHash (15.03%), VAEH (14.74%), BihalfHash (17.36%) to 17.86%. Thus, compared with other state-of-art retrieval approaches, the UTBH approach achieves better performance, which demonstrates the effectiveness of the proposed UTBH approach.

V. CONCLUSION

In this article, we develop a novel unsupervised multispectral RS image retrieval approach, which learns hash codes by guiding the interactions among the multichannel information of multispectral RS images and reducing the disparity between the hash-like codes space and the Hamming space adequately. Firstly, the proposed UTBH method utilizes the transformer for solving the issue of deficient usage of the interactions among the multichannel information of multispectral RS images. Second, we perform effective hash codes learning by designing a novel objective function, which not only preserves discrimination of hash codes in the learning process but also reduce the disparity between the hash-like codes space and the Hamming space. Finally, comprehensive experiments on diverse benchmark have well demonstrated the reasonableness and effectiveness of the proposed UTBH method.

REFERENCES

[1] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing Big Data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, 2021.

[2] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4076–4084.

[3] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.

[4] H. Ning, X. Zheng, Y. Yuan, and X. Lu, "Audio description from image by modal translation network," *Neurocomputing*, vol. 423, pp. 124–134, 2021.

[5] L. Peng and R. Peng, "Partial randomness hashing for large-scale remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 464–468, Mar. 2017.

[6] Y. Li, Y. Zhang, H. Xin, Z. Hu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 950–965, Feb. 2018.

[7] Y. Ma et al., "Remote sensing Big Data computing: Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 51, pp. 47–60, 2015.

[8] M. Zhang, Q. Cheng, F. Luo, and L. Ye, "A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2711–2723, 2021.

[9] H. Zhao, L. Yuan, H. Zhao, and Z. Wang, "Global-aware ranking deep metric learning for remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8008505.

[10] Y. Zhan, R. Osolo Ian, S. Wuqing, and L. Jun, "Deep attention-guided hashing," *IEEE Access*, vol. 7, pp. 11209–11221, 2019.

[11] H. Wang, B. Qu, X. Lu, and Y. Chen, "Unsupervised variational autoencoder hash algorithm based on multi-channel feature fusion," in *Proc. 12th Int. Conf. Digit. Image Process.*, 2020, pp. 433–440.

[12] P. Li, X. Zhang, X. Zhu, and P. Ren, "Online hashing for scalable remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 709.

[13] B. Zhu, M. Ramsey, and H. Chen, "Creating a large-scale content-based airphoto image digital library," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 163–167, Jan. 2000.

[14] J. Li and R. M. Narayanan, "Integrated spectral and spatial information mining in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 673–685, Mar. 2004.

[15] G. J. Scott, M. N. Klaric, C. H. Davis, and C.-R. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1603–1616, May 2011.

[16] Z. Shao, W. Zhou, L. Zhang, and J. Hou, "Improved color texture descriptors for remote sensing image retrieval," *J. Appl. Remote Sens.*, vol. 8, no. 1, 2014, Art. no. 083584.

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

[18] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–8.

[19] M. An, Z. Jiang, and D. Zhao, "High speed robust image registration and localization using optimized algorithm and its performances evaluation," *J. Syst. Eng. Electron.*, vol. 21, no. 3, pp. 520–526, 2010.

[20] S. Newsam and Y. Yang, "Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery," in *Proc. 15th Annu. ACM Int. Symp. Adv. Geograph. Inf. Syst.*, 2007, pp. 1–8.

[21] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2012.

[22] W. Zhou, Z. Shao, C. Diao, and Q. Cheng, "High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder," *Remote Sens. Lett.*, vol. 6, no. 10, pp. 775–783, 2015.

[23] Y. Chen, S. Xiong, L. Mou, and X. X. Zhu, "Deep quadruple-based hashing for remote sensing image-sound retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4705814.

[24] F. Ye, Y. Su, H. Xiao, X. Zhao, and W. Min, "Remote sensing image registration using convolutional neural network features," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 232–236, Feb. 2018.

[25] Y. Liu, L. Ding, C. Chen, and Y. Liu, "Similarity-based unsupervised deep transfer learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7872–7889, Nov. 2020.

[26] N. S. Kumar, M. Arun, and M. K. Dangi, "Remote sensing image retrieval using object-based, semantic classifier techniques," *Int. J. Inf. Commun. Technol.*, vol. 13, no. 1, pp. 68–82, 2018.

[27] R. Imbriaco et al., "Aggregated deep local features for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 493.

[28] D. Ye, Y. Li, C. Tao, X. Xie, and X. Wang, "Multiple feature hashing learning for large-scale remote sensing image retrieval," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 11, 2017, Art. no. 364.

[29] T. Reato, B. Demir, and L. Bruzzone, "An unsupervised multicode hashing method for accurate and scalable remote sensing image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 276–280, Feb. 2019.

[30] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, and A. Plaza, "Deep hashing based on class-discriminated neighborhood embedding," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5998–6007, 2020.

[31] L. Han, P. Li, X. Bai, C. Grecos, X. Zhang, and P. Ren, "Cohesion intensive deep hashing for remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 101.

[32] X. Lu, Y. Chen, and X. Li, “Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 106–120, Jan. 2018.

[33] Y. Chen and X. Lu, “Deep discrete hashing with pairwise correlation learning,” *Neurocomputing*, vol. 385, pp. 111–121, 2020.

[34] X. Lu, Y. Chen, and X. Li, “Discrete deep hashing with ranking optimization for image retrieval,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2052–2063, Jun. 2020.

[35] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 448–456.

[36] Y. Zhen, J. W. Keung, Y. Xiao, X. Yan, J. Zhi, and J. Zhang, “On the significance of category prediction for code-comment synchronization,” *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 2, 2022, Art. no. 30.

[37] Y. Zhang, X. Zheng, and X. Lu, “Remote sensing cross-modal retrieval by deep image-voice hashing,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9327–9338, 2022.

[38] X. Ma, J. Keung, Z. Yang, X. Yu, Y. Li, and H. Zhang, “CASMS: Combining clustering with attention semantic model for identifying security bug reports,” *Inf. Softw. Technol.*, vol. 147, 2022, Art. no. 106906.

[39] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[40] Z. Yang et al., “A multi-modal transformer-based code summarization approach for smart contracts,” in *Proc. IEEE/ACM 29th Int. Conf. Prog. Comprehension*, 2021, pp. 1–12.

[41] Y. Chen, H. Dai, X. Yu, W. Hu, Z. Xie, and C. Tan, “Improving ponzi scheme contract detection using multi-channel textCNN and transformer,” *Sensors*, vol. 21, no. 19, 2021, Art. no. 6417.

[42] A. Vaswani et al., “Attention is all you need,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[43] H. Liu, R. Wang, S. Shan, and X. Chen, “Deep supervised hashing for fast image retrieval,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2064–2072.

[44] Y. Weiss et al., “Spectral hashing,” in *Proc. 21st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.

[45] T.-T. Do, A.-D. Doan, and N.-M. Cheung, “Learning to hash with binary deep neural network,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 219–234.

[46] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, “Deepsat: A learning framework for satellite imagery,” in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2015, pp. 1–10.

[47] G.-S. Xia et al., “AID: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[48] A. Shukla, A. Upadhyay, M. Sharma, V. Chinnusamy, and S. Kumar, “High-resolution NIR prediction from RGB images: Application to plant phenotyping,” in *Proc. IEEE Int. Conf. Image Process.*2022, pp. 4058–4062.

[49] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, “Investigation on the stability of smote-based oversampling techniques in software defect prediction,” *Inf. Softw. Technol.*, vol. 139, 2021, Art. no. 106662.

[50] Y. Chen, X. Lu, and X. Li, “Supervised deep hashing with a joint deep network,” *Pattern Recognit.*, vol. 105, 2020, Art. no. 107368.

[51] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[52] Y. Chen and X. Lu, “Deep category-level and regularized hashing with global semantic similarity learning,” *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6240–6252, Dec. 2021.

[53] F. Shen, C. Shen, Q. Shi, A. Hengel, and Z. Tang, “Inductive hashing on manifolds,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1562–1569.

[54] W. Kong and W. Li, “Isotropic hashing,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1655–1663.

[55] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, “Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[56] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon, “Spherical hashing: Binary code embedding with hyperspheres,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2304–2316, Nov. 2015.

[57] B. Demir and L. Bruzzone, “Hashing-based scalable remote sensing image search and retrieval in large archives,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 892–904, Feb. 2016.

[58] L.-K. Huang, Q. Yang, and W.-S. Zheng, “Online hashing,” in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1422–1428.

[59] C. Leng, J. Wu, J. Cheng, X. Bai, and H. Lu, “Online sketching hashing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2503–2511.

[60] S. Su, C. Zhang, K. Han, and Y. Tian, “Greedy hash: Towards fast optimization for accurate hash coding in CNN,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 806–815.

[61] Y. Li and J. Van Gemert, “Deep unsupervised image hashing by maximizing bit entropy,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 2002–2010.

Yaxiong Chen received the B.Sc. degree in mathematics from Hubei University, China, in 2014, the M.Sc. degree in mathematics from Wuhan University of Technology, China, in 2017, and the Ph.D. degree in signal and information processing from University of Chinese Academy of Sciences, China, in 2020.

He is an Associate Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China. His current research interests include pattern recognition, machine learning, hyperspectral image analysis, and medical imaging.

Fan Wang received the master’s of engineering degree in computer science and technology from the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China, in 2023.

His main research interests include image retrieval and style transfer.

Lin Lu received the M.Sc. and B.Sc. degrees in computer science and technology from Yangzhou University, Yangzhou, China, in 2011 and 2014, respectively. He is currently working toward the Ph.D. degree with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China, in 2023.

His research interests include intelligent connected vehicles, big data, and machine learning.

Shengwu Xiong received the B.Sc. degree in computational mathematics and M.Sc. and Ph.D. degrees in computer software and theory from Wuhan University, Wuhan, China, in 1987, 1997, and 2003, respectively.

He is currently a Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, China. His research interests include intelligent computing, machine learning, and pattern recognition.